SINGLE COPY GENOMIC HYBRIDIZATION PROBES AND METHOD OF GENERATING SAME

RELATED APPLICATION

This is a continuation-in-part of Serial No. 09/573,080 filed May 16, 2000.

SEQUENCE LISTING

A Sequence Listing containing 613 sequences in the form of a computer readable ASCII file in connection with the present invention is incorporated herein by reference and appended hereto as one (1) original compact disk in accordance with 37 CFR 1.821(c), an identical copy thereof in accordance with 37 CFR 1.821(e), and one (1) identical copy thereof in accordance with 37 CFR 1.52(e).

COMPUTER PROGRAM LISTING APPENDIX

A computer program listing appendix containing the source code of a computer program that may be used with the present invention is incorporated herein by reference and appended hereto as one (1) original compact disk, and an identical copy thereof, containing a total of 3 files as follows:

Date of Creation	Size (Bytes)	File Name
04/18/01	26 KB	FINDI.PL
04/18/01	19 KB	PRIM.IN
04/18/01	20 KB	PRIM.WKG

BACKGROUND OF THE INVENTION

Field of the Invention

The present invention is broadly concerned with a method for designing single copy hybridization probes useful in the fields of cytogenetics and molecular genetics for determining the presence of specific nucleic acid sequences in a sample of eukaryotic origin, e.g., the probes may be used to analyze specific chromosomal locations by *in situ* hybridization as a detection of acquired or inherited genetic diseases. More particularly, the

invention pertains to such probes, hybridization methods of use thereof and techniques for developing the probes, where the probes are essentially free of genomic repeat sequences, thereby eliminating the need for disabling of repetitive sequences which is required with conventional probes.

Description of the Prior Art

Chromosome abnormalities are associated with various genetic disorders, which may be inherited or acquired. These abnormalities are of three general types, extra or missing individual chromosomes (aneuploidy), extra or missing portions of chromosomes (including deletions, duplications, supernumerary and marker chromosomes), or chromosomal rearrangements. The latter category includes translocations (transfer of a piece from one chromosome onto another chromosome), inversions (reversal in polarity of a chromosomal segment), insertions (transfer of a piece from one chromosome into another chromosome) and isochromosomes (chromosome arms derived from identical chromosomal segments). The abnormalities may be present only in a subset of cells (mosaicism), or in all cells. Inherited or constitutional abnormalities of various types occur with a frequency of about one in every 250 human births, with results which may be essentially benign, serious or even lethal. Chromosomal abnormalities are common and often diagnostic in acquired disorders such as leukemia and other cancers.

Hybridization probes have been developed in the past for chromosome analysis and diagnosis of abnormalities. The probes comprise cloned or amplified genomic sequences or cDNA. For example, U.S. Patents Nos. 5,447,841, 5,663,319 and 5,756,696 describe hybridization probes in the form of labeled nucleic acids which are complementary to nucleic acid segments within target chromosomal DNA. However, these probes contain repetitive sequences and therefore must be used in conjunction with blocking nucleic acids which are substantially complementary to repetitive sequences in the labeled probes. That is, these prior art probes are either pre-reacted with the blocking nucleic acids so as to bind and block the repetitive sequences therein, or such blocking nucleic acids are present in the hybridization reaction mixture. If the repetitive sequences in the probes are not disabled in some manner, the probes will react with the multiple locations in the target chromosomal DNA where the repetitive sequences reside and will not specifically react with the single

copy target sequences. This problem is particularly acute with interspersed repeat sequences which are widely scattered throughout the genome, but also is present with tandem repeats clustered or contiguous on the DNA molecule. The requirement for repeat sequence disabilization by using complementary blocking nucleic acids reduces the sensitivity of the existing probes. Reliable, easily detectable signals require DNA probes of from about 40-100 kb.

The prior art also teaches that cloned probes presumed to contain single copy sequences can be identified based on their lack of hybridization to radiolabeled total genomic DNA. In these other studies, hybridization is first performed with probes that contain pools of clones in which each recombinant DNA clone has been individually selected so that it hybridizes to single-copy sequences or very low copy repetitive sequences. A prerequisite step in this prior art is to identify single copy sequences by experimental hybridization of labeled genomic DNA to a candidate DNA probe by Southern or dot-blot hybridization. Positive hybridization with labeled total genomic DNA usually indicates that the candidate DNA probe contains a repetitive sequence and eliminates it from consideration as a single copy probe. Furthermore, an experimental hybridization of a DNA probe with total genomic DNA may fail to reveal the presence of multicopy repetitive sequences that are not abundant (<100 copies) or are infrequent in the genome. Such sequences represent a small fraction of the labeled genomic DNA and the signal they contribute will be below the limits of detection.

It has also been suggested to physically remove repeat sequences from probes by experimental procedures (Craig et al., *Hum. Genet.*, **100:**472-476 (1997); Durm et al., *Biotech.*, **24:**820-825 (1998)). This procedure involves prehybridizing a polymerase chain reaction (PCR)-amplified genomic probe with an excess of purified repetitive sequence DNA prior to applying the probe to the DNA target. The resulting purified probe is depleted of repetitive sequences. This procedure is in principle very similar to other procedures that disable the hybridization of repetitive sequences in probes, but the technique is time-consuming and does not provide any advantages over the probes described in Patents Nos. 5,447,841 and 5,756,696.

SUMMARY OF THE INVENTION

The present invention overcomes the problem outlined above and provides nucleic acid (e.g., DNA) hybridization probes comprising a labeled, single copy nucleic acid which hybridizes with a deduced single copy sequence interval in target nucleic acid of known sequence. Generally speaking, the probes of the invention are designed by comparing the sequence of a target nucleic acid with known repeat sequences in the genome of which the target is a part; with this information it is possible to deduce the single copy sequences within the target (i.e., those sequences which are essentially free of repeat sequences which, due to the lack of specificity, can mask the hybridization signal of the single copy sequences). As can be appreciated, these initial steps require knowledge of the sequences both of the target and genomic repeats, information which is increasingly available owing to the Human Genome Project and related bioinformatic studies. Furthermore, readily available computer software is used to derive the necessary single copy sequences.

The probes hereof are most preferably complementary to the target sequence, i.e., there is a 100% complementary match between the probe nucleotides and the target sequence. More broadly, less than 100% correspondence probes can be used, so long as the probes adequately hybridize to the target sequence, i.e., there should be at least about 80% sequence identity between the probe and a sequence which is a complement to target sequences, more preferably at least about 90% sequence identity.

Nucleic acid fragments corresponding to the deduced single copy sequences can be generated by a variety of methods, such as PCR amplification, restriction or exonuclease digestion of purified genomic fragments, or direct nucleic acid synthesis. The single copy fragments are then purified to remove any potentially contaminating repeat sequences, such as, for example, by electrophoresis or denaturing high pressure liquid chromatography; this is highly desirable because it eliminates spurious hybridization and detection of unrelated genomic sequences.

The probe fragments may then be cloned into a recombinant DNA vector or directly labeled. The probe is preferably labeled by nick translation using a modified or directly labeled nucleotide. The labeled probe is then denatured and hybridized, preferably to fixed chromosomal preparations on microscope slides or alternately to purified nucleic acid immobilized on a filter, slide, DNA chip, or other substrate. The probes can then be

hybridized to chromosomes according to conventional fluorescence *in situ* hybridization (FISH) methods such as those described in U.S. Patents Nos. 5,985,549 or 5,447,841; alternately, they can be hybridized to immobilized nucleic acids according to the techniques described in Patents Nos. 5,110,920 or 5,273,881. Probe signals may be visualized by any of a variety of methods, such as those employing fluorescent, immunological or enzymatic detection reagents.

Use of the probes of the invention permits more precise chromosomal breakpoint determinations, to a level of resolution heretofore unobtainable by *in situ* hybridization. In such analyses, initial probe sets can be prepared from regions believed to be on opposite sides of the breakpoint. After an initial assay to confirm this, successive additional probes closer to the breakpoint can be designed, using the single copy strategy. In this fashion, the precise region of the breakpoint can be determined.

It has been found that use of putative single copy probes can determine the existence of heretofore unknown repeat sequences in a genome. The heretofore unknown repeated sequence families can then be included in the repetitive sequence database so that these sequences can be used in the design of subsequent single copy probes.

It was also found that probes may contain sequences that are duplicated or triplicated in the genome which can have stronger hybridization due to the increased length of the target sequence. Also, these duplicons or triplicons can be confirmed, as such, using single copy probes which is more difficult with available commercial probes.

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1-12 are respective CCD camera images of FISH experiments wherein various gene-specific digoxigenin-dUTP labeled probes were hybridized on metaphase cells and detected with rhodamine conjugated antibody to digoxigenin and where the chromosomes were counterstained with 4,6-diamidino-2-phenylindole (DAPI). Chromosomes with one or both chromatids hybridized are indicated by arrows whereas a star indicates the absence of normally expected hybridizations. In particular,

Fig. 1 illustrates hybridization results using the 5170 bp HIRA probe described in Example 1, and wherein the probe was reacted with purified repetitive DNA sequences;

Fig. 2 illustrates a comparative hybridization identical to that depicted in Fig. 1, using the same 5170 bp HIRA probe but without pre-reaction with purified repetitive DNA sequences;

Fig. 3 illustrates hybridization results using the 3544 bp 15q11-q13 probe pre-reacted with purified repetitive DNA;

Fig. 4 illustrates results in a comparative experiment using the 3544 bp 15q11-q13 probe without pre-reaction with purified repetitive DNA;

Fig. 5 illustrates hybridization results using the 4166 bp, 3544 bp and 2290 bp 15q11-q13 probes described in Example 2, without pre-reaction with purified repetitive DNA sequences;

Fig. 6 illustrates hybridization results using the 5170 bp, 3691 bp, 3344 bp and 2848 bp HIRA probes described in Example 1 without pre-reaction with purified repetitive DNA sequences;

Fig. 7 illustrates hybridization results using the 4823 bp 1p36.3 probe described in Example 2 on metaphase cells of a normal individual, with pre-reaction with purified repetitive DNA sequences;

Fig. 8 illustrates a comparative hybridization result using the 4823 bp 1p36.3 probe of Fig. 7 without pre-reaction with purified repetitive DNA sequences;

Fig. 9 illustrates hybridization results using the 4724 bp and 4823 bp 1p36.3 probes described in Example 2 with pre-reaction with purified repetitive DNA sequences, and wherein single copy hybridizations were observed on homologous pairs of chromosome 1s;

Fig. 10 illustrates a comparative hybridization result using the 4724 bp and 4823 bp 1p36.3 probes described in Example 2 without pre-reaction with purified repetitive DNA sequences, and depicting the same single copy hybridizations shown in Fig. 9;

Fig. 11 illustrates hybridization results using the 4166 bp, 3544 bp and 2290 bp 15q11-q13 probes described in Example 2 without pre-reaction with purified DNA sequences on metaphase cells of a patient affected with Prader-Willi syndrome and known to harbor a deletion of 15q11-q13 sequences for one chromosomal allele, with a star indicating lack of hybridization at the deleted chromosome position and with the arrow indicating hybridization to a single chromosome;

Fig. 12 illustrates hybridization results using the 3691 bp, 3344 bp and 2848 bp HIRA probes described in Example 1 without pre-reaction with purified DNA sequences on metaphase cells of a patient affected with DiGeorge/Velo-Cardio-Facial Syndrome (VCFS) known to harbor a deletion of 22q11.2 sequences, wherein the star indicating lack of hybridization at the deleted chromosome position and the arrow indicating a normal homolog;

Fig. 13 is a scatterplot of base pair coordinates versus single copy probe lengths found in the Breakage Cluster Region gene (BCR) promoter found on chromosome 22, the disruption of which is common in cases of chronic adult myeloid leukemia and in some cases of acute lymphoblastic leukemia, as described in Example 4;

Fig. 14 is a scatterplot of base pair coordinates versus single copy probe lengths found in the ABL1 gene on chromosome 9, the disruption of which is common in cases of chronic adult myeloid leukemia and in some cases of acute lymphoblastic leukemia, as described in Example 4;

Fig. 15 is a CCD camera image of a FISH experiment using chromosome 9q34-specific, digoxigenin-dUTP labeled probes from the ABL1 oncogene (SEQ ID Nos. 520-525), and detected with rhodamine (red) conjugated antibody to digoxigenin with DAPI stained metaphase cells from a patient with chronic myelogenous leukemia (CML), illustrating the use of probes downstream of the site of fusion between ABL1 gene and BCR gene used to make a precise chromosomal breakpoint determination as explained in Example 4 wherein the derivative chromosome 22 and normal chromosome 9 are indicated; and

Fig. 16 is a CCD camera image of a FISH experiment using chromosome 9q34-specific, digoxigenin-dUTP labeled probes from the ABL1 oncogene (SEQ ID Nos. 516-525) and detected with rhodamine (red) conjugated antibody to digoxigenin, with DAPI stained metaphase cells from a patient with CML, illustrating the use of probes from each side of the site of fusion between BCR and ABL1 genes used to make a precise chromosomal 9q34 breakpoint determination as explained in Example 4, wherein the derivative chromosome 22, derivative chromosome 9 and normal chromosome 9 are indicated.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENT

The present invention is concerned with nucleic acid (e.g., DNA) hybridization probes useful for detection of genetic or neoplastic disorders. The probes are in the form of labeled nucleic acid fragments or a collection of labeled nucleic acid fragments whose hybridization to a target sequence can be detected. The invention also pertains to methods of developing, generating and labeling such probes, and to uses thereof.

The labeled probes hereof may be used with any nucleic acid target that may potentially contain repetitive sequences. These target sequences may include, but are not limited to chromosomal or purified nuclear DNA, heteronuclear RNA, or mRNA species that contain repetitive sequences as integral components of the transcript. In the ensuing detailed explanation, the usual case of a DNA target sequence and DNA probes is discussed; however, those skilled in the art will understand that the discussion is equally applicable (with art-recognized differences owing to the nature of the target sequences and probes) to other nucleic acid species.

An important characteristic of the probes of the invention is that they are composed of "single copy" or "unique" DNA sequences which are both complementary to at least a portion of the target DNA region of interest and are essentially free of sequences complementary to repeat sequences within the genome of which the target region is a part. Accordingly, a probe made up of a single copy or unique sequence is complementary to essentially only one sequence in the corresponding genome.

Very recently, it has been discovered that the human genome contains highly similar domains which have been termed duplicons when they are present in two non-allelic copies or triplicons when present in three copies in the genome (Ji et al., *Genome Res.*, **10**:597-610 (2000)). Duplication or triplication of chromosomal domains containing such sequences were recent evolutionary events, based on the fact that non-human primates, in some instances, do not contain multiple copies of these sequences, and on the high degree of sequence similarity between different copies of paralogous sequences. These low copy duplicons (or triplicons) are to be distinguished from classic repetitive sequence families, which tend to either be interspersed throughout the genome or to be tandemly reiterated hundreds to thousands of times in the same chromosomal interval; therefore, probes from

duplicons or triplicons are for purposes of the present invention deemed to be within the ambit of "single copy" probes.

These duplicons or triplicons have evolved so recently that the sequence and organization of an entire genomic domain – which comprises complex, near-single copy segments and adjacent members of known repetitive sequence families – is completely conserved in each duplicon or triplicon segment. Duplicon and triplicon lengths of several kilobases to megabase sizes have been reported (International Genome Sequencing Consortium, *Nature*, **409**:860-922 (2001)). The duplicons/triplicons are often tandemly arranged, and are almost always present on the same chromosome, and are therefore clustered in the genome. The length of the interval separating paralogous probe sequences is dictated by the size of the duplicated/triplicated domain, the orientation of the duplicons (or triplicons) relative to each other (ie. direct or inverted), and the length of unrelated sequence intervals, if any, which separate the duplicons/triplicons.

In the context of the present invention, the term "single copy" with reference to a nucleic acid sequence refers to a sequence which is strictly unique (i.e., which is complementary to one and one only sequence in the corresponding genome) but also covers duplicons and triplicons. Stated otherwise, a "single copy" probe in preferred forms will hybridize to three or less locations in the genome.

As used herein, a "repeat sequence" is a sequence which repeatedly appears in the genome of which the target DNA is a part, with a sequence identity between repeats of at least about 60%, more preferably at least about 80%, and which is of sufficient length or has other qualities which would cause it to interfere with the desired specific hybridization of the probe to the target DNA (i.e., the probe would hybridize with multiple copies of the repeat sequence). Generally speaking, a repeat sequence appears at least about 10 times in the genome (more preferably at least about 50 times, and most preferably at least about 200 times) and has a length of at least about 50 nucleotides, and more preferably at least about 100 nucleotides. Repeat sequences can be of any variety (e.g., tandem, interspersed, palindromic or shared repetitive sequences with some copies in the target region and some elsewhere in the genome), and can appear near the centromeres of chromosomes, distributed over a single chromosome, or throughout some or all chromosomes. Normally, with but few exceptions, repeat sequences do not express physiologically useful proteins.

Repetitive sequences occur in multiple copies in the haploid genome. The number of copies can range from at least about 10 to hundreds of thousands, wherein the Alu family of repetitive DNA are exemplary of the latter numerous variety. The copies of a repeat may be clustered or interspersed throughout the genome. Repeats may be clustered in one or more locations in the genome, for example, repetitive sequences occurring near the centromeres of each chromosome, and variable number tandem repeats (VNTRs) (Nakamura et al., *Science*, **235**:1616 (1987)); or the repeats may be distributed over a single chromosome for example, repeats found only on the X chromosome as described by Bardoni et al. (*Cytogenet. Cell Genet.*, **46**:575 (1987)); or the repeats may be distributed over all the chromosomes, for example, the Alu family of repetitive sequences.

Simple repeats of low complexity can be found within genes but are more commonly found in non-coding genomic sequences. Such repeated elements consist of mono-, di-, tri-, tetra-, or penta-nucleotide core sequence elements arrayed in tandem units. Often the number of tandem units comprising these repeated sequences varies at the identical locations among genomes from different individuals. These repetitive elements can be found by searching for consecutive runs of the core sequence elements in genomic sequences.

As used herein, "sequence identity" refers to a relationship between two or more polynucleotide sequences, namely a reference sequence and a given sequence to be compared with the reference sequence. Sequence identity is determined by comparing the given sequence to the reference sequence after the sequences have been optimally aligned to produce the highest degree of sequence similarity, as determined by the match between strings of such sequences. Upon such alignment, sequence identity is ascertained on a position-by-position basis, e.g., the sequences are "identical" at a particular position if at that position, the nucleotides are identical. The total number of such position identities is then divided by the total number of nucleotides or residues in the reference sequence to give % sequence identity. Sequence identity can be readily calculated by known methods, including but not limited to, those described in Computational Molecular Biology, Lesk A. N., ed., Oxford University Press, New York (1988); Biocomputing: Informatics and Genome Projects, Smith D.W., ed., Academic Press, New York (1993); Computer Analysis of Sequence Data, Part I, Griffin A.M., and Griffin H. G., eds., Humana Press, New Jersey (1994); Sequence Analysis in Molecular Biology, von Heinge G., Academic Press (1987);

Sequence Analysis Primer, Gribskov M. and Devereux J., eds., M. Stockton Press, New York (1991); and Carillo H., and Lipman D., SIAM J. Applied Math., **48**:1073 (1988). Preferred methods to determine the sequence identity are designed to give the largest match between the sequences tested. Methods to determine sequence identity are codified in publicly available computer programs which determine sequence identity between given sequences. Examples of such programs include, but are not limited to, the GCG program package (Devereux et al., *Nuc. Ac. Res.*, **12**(1):387 (1984)), BLASTP, BLASTN and FASTA (Altschul et al., *J. Molec. Biol.*, **215**:403-410 (1990)). The BLASTX program is publicly available from NCBI and other sources (BLAST Manual, Altschul et al., NCBI, NLM, NIH, Bethesda, MD 20894; Altschul et al., *J. Molec. Biol.*, **215**:403-410 (1990)). These programs optimally align sequences using default gap weights in order to produce the highest level of sequence identity between the given and reference sequences. As an illustration, by a polynucleotide having a nucleotide sequence having at least, for example, 95% "sequence identity" to a reference nucleotide sequence, it is intended that the nucleotide sequence of the given polynucleotide is identical to the reference sequence except that the given polynucleotide sequence may include up to 5 differences per each 100 nucleotides of the reference nucleotide sequence. In other words, in a polynucleotide having a nucleotide sequence having at least 95% identity relative to the reference nucleotide sequence, up to 5% of the nucleotides in the reference sequence may be deleted or substituted with another nucleotide, or a number of nucleotides up to 5% of the total nucleotides in the reference sequence may be inserted into the reference sequence. Inversions in either sequence are detected by these computer programs based on the similarity of the reference sequence to the antisense strand of the homologous test sequence. These variants of the reference sequence may occur at the 5' or 3' terminal positions of the reference nucleotide sequence or anywhere between those terminal positions, interspersed either individually among nucleotides in the reference sequence or in one or more contiguous groups within the reference sequence.

The single copy probes of the invention preferably should have a length of at least about 50 nucleotides, and more preferably at least about 100 nucleotides. Probes of this length are sufficient for Southern blot analyses. However, if other analyses such as FISH are employed, the probes should be somewhat longer, i.e., at least about 500 nucleotides, and

more preferably at least about 2000 nucleotides in length. The probes can be used to detect virtually any type of chromosomal rearrangement, such as deletions, duplications, insertions, additions, inversions or translocations.

In order to develop probes in accordance with the invention, the sequence of the target DNA region must be known. The target region may be an entire chromosome or only portions thereof where rearrangements have been identified. With this sequence knowledge, the objective is to determine the boundaries of single copy or unique sequences within the target region. This is preferably accomplished by inference from the locations of repetitive sequences within the target region. Normally, the sequence of the target region is compared with known repeat sequences from the corresponding genome, using available computer software. Once the repeat sequences within the target region are identified, the intervening sequences are deduced to be single copy (i.e., the sequences between adjacent repeat sequences).

Optimal alignment of the target and repetitive sequences for comparison may be conducted by the local homology algorithm of Smith et al., *Adv. Appl. Math.*, **2:**482 (1981), by the homology alignment algorithm of Needleman et al., *J. Mol. Biol.*, **48:**443 (1970). The results obtained from the heuristic methods (Pearson et al., *Proc. Natl. Acad. Sci.*, **85:**244 (1988); Altschul et al., *J. Molec. Biol.*, **215**:403-410 (1990)) are generally not as comprehensive as the methods of Smith et al. (1981) and Needleman et al., (1970). However, they are faster than these methods.

Once the single copy sequence information is obtained, certain of the single copy sequences (normally the longest) are used to design hybridization probes. In this regard, probes may be of varying "complexity" as defined by Britten et al., *Methods of Enzymol.*, **29:**363 (1974) and as further explained by Cantor et al., *Biophysical Chemistry: Part III: The Behavior of Biological Macromolecules,* pp. 1228-1230. The complexity of selected probes is dependent upon the application for which it is designed. In general, the larger the target area, the more complex the probe. The complexity of a probe needed to detect a set of sequences will decrease as hybridization sensitivity increases. At high sensitivity and low background, smaller and less complex probes can be used.

With current hybridization techniques, it is possible to obtain reliable, easily detectable signals with relatively small probes in accordance with the invention. A readily

detectable signal was obtained with a probe on the order of 2 kb in length, using FISH technology. This sensitivity of the present method is improved compared to the prior art (U.S. Patent No. 5,756,696) because the probes of the present invention are homogeneous single copy sequences. However, smaller amplified segments, each comprising non-repetitive sequences, may also be used in combination as probes to achieve adequate signals for *in situ* hybridization. Complex single copy probes that hybridize to duplicated or triplicated targets can also increase hybridization signals.

One application of the use of multiple fragment probes is in the detection of translocations between different chromosomes. Proportionately increasing the complexity of the probe also permits analysis of multiple compact regions of the genome simultaneously. For a single chromosome, the portion of the probe targeted to one side of the breakpoint can be labeled and detected differently from that targeted to the other side of the breakpoint so that the derivative or translocated chromosome is detected by one label and is distinguishable from the intact normal chromosome which has both labels.

The invention makes it possible to produce single copy probes at a higher genomic density than possible using conventional probes. Chromosomes 21 and 22 have been comprehensively sequenced, and it has been determined that adjacent single copy intervals tend to be clustered on these chromosomes. For example, on chromosome 22, 39% of single copy intervals are separated by only 500-1000 bp. Single copy intervals ≥2.3 kb are separated, on average, by 29.2 kb on chromosome 21 and by 22.3 kb on chromosome 22.

In order to estimate the size of genomic intervals required to develop single copy probes, the probability of detecting at least one single copy sequence in overlapping, uniform-length genomic intervals on chromosomes 21q and 22q was determined. Single copy segments ≥2.0 kb in length are found in the majority of 100 kb genomic intervals of these chromosomes (96% of chromosome 22q and 88% of chromosome 21q). Increasing the size of the genomic sequence to 150 kb results in 99% coverage of chromosome 22q and 96% of chromosome 21q. Therefore, single copy probes should be more or less 2 kb to ensure comprehensive coverage (at least once per 100-150 kb) of chromosomes 21 and 22. Assuming that single copy sequences are similarly distributed on other chromosomes, it should be feasible to develop probes for *in situ* hybridization analysis of most clinically relevant chromosomal rearrangements.

Once appropriate single copy sequences in the chromosomal region of interest have been identified, PCR is preferably used for amplifying the appropriate DNA to obtain probes. PCR is a well known technique for amplifying specific DNA segments in geometric progression and relies upon repeated cycles of DNA polymerase-catalyzed extension from a pair of oligonucleotide primers with homology to the 5' end and to the complement of the 3' end of the DNA segment to be amplified.

The nucleic acid (e.g., DNA) that serves as the PCR template may be single stranded or double stranded, but when the DNA is single stranded, it will typically be converted to double stranded. The length of the template DNA may be as short as 50 bp, but usually will be at least about 100 bp long, and more usually at least about 150 bp long, and may be as long as 10,000 bp or longer, but will usually not exceed 50,000 bp in length, and more usually will not exceed 20,000 bp in length. The DNA may be free in solution, flanked at one or both ends with non-template DNA, present in a cloning vector such as a plasmid and the like, with the only criteria being that the DNA be available for participation in the primer extension reaction. The template DNA may be derived from a variety of different sources, so long as it is complementary to the target chromosomal or immobilized DNA sequence. The amount of template DNA that is combined with the other reagents will range from about 1 molecule to 1 pmol, usually from about 50 molecules to 0.1 pmol, and more usually from about 0.01 pmol to 100 fmol. The oligonucleotide primers with which the template nucleic acid is contacted will be of sufficient length to provide for hybridization to complementary template DNA under annealing conditions but will be of insufficient length to form stable hybrids with template DNA under polymerization conditions. The primers will generally be at least about 10 nucleotides (nt) in length, usually at least 15 nt in length and more usually at least 16 nt in length and may be as long as 30 nt in length or longer, where the length of the primers will generally range from 18 to 50 nt in length, usually from about 20 to 35 nt in length. The yield of longer amplification products can be enhanced using primers of 30 to 35 nt and high fidelity polymerases (described in U.S. Patent No. 5,436,149).

To maximize the signal intensity obtained during *in situ* hybridization, primer sequence pairs are preferred which, upon amplification, produce a DNA fragment that spans nearly the entire length of each single-copy genomic sequence interval. Hence, contiguous or closely spaced (software excludes pairs that are separated by ≤70% of the length of the

single copy interval) primer pairs are generally excluded from consideration for producing probes for *in situ* hybridization. With the exception of cytogenetic preparations, this criterion is generally not applicable for probes that are hybridized to immobilized cloned or synthetic nucleic acid targets, since signal intensities of shorter probes are usually adequate due to the increased number of target molecules.

However, in order to optimize the yield and kinetics of the PCR reaction, the desired primer sequences are also subject to other criteria. First, a primer sequence should not be substantially self-complementary or complementary to the second primer. In particular, potential primer sequences are excluded which could result in the formation of stable hybrids involving the 3' terminus of the primer and either another sequence in the same or the second primer (defined as ≥6 base pairs). Additionally, the T_m of one member of the primer pair should occur within 2°C of its counterpart, which enables them to denature and anneal to the template nearly simultaneously. Software is well known in the art to identify primer sequences that satisfy all of the preferred criteria (see for example: http://www-genome.wi.mit.edu/ftp/pub/software/primer.0.5/ or http://www.oligo.net/Oligo_6_tour.htm).

The PCR reaction mixture will normally further comprise an aqueous buffer medium which includes a source of monovalent ions, a source of divalent cations and a buffering agent. Any convenient source of monovalent ions, such as KCl, K-acetate, NH_4-acetate, K-glutamate, NH_4 Cl, ammonium sulfate, and the like may be employed, where the amount of monovalent ion source present in the buffer will typically be present in an amount sufficient to provide for a conductivity in a range from about 500 to 20,000, usually from about 1000 to 10,000, and more usually from about 3,000 to 6,000 microohms. The divalent cation may be magnesium, manganese, zinc and the like, where the cation will typically be magnesium. Any convenient source of magnesium cation may be employed, including $MgCl_2$, Mg-acetate, and the like. The amount of Mg^{+2}- present in the buffer may range from 0.5 to 10 mM, but will preferably range from about 2 to 4 mM, more preferably from about 2.25 to 2.75 mM and will ideally be at about 2.45 mM. Representative buffering agents or salts that may be present in the buffer include Tris, Tricine, HEPES, MOPS and the like, where the amount of buffering agent will typically range from about 5 to 150 mM, usually from about 10 to 100 mM, and more usually from about 20 to 50 mM, where in certain preferred

embodiments the buffering agent will be present in an amount sufficient to provide a pH ranging from about 6.0 to 9.5, where most preferred is pH 7.3 at 72°C. Other agents which may be present in the buffer medium include chelating agents, such as EDTA, EGTA and the like.

Also present in the PCR reaction mixtures is a melting point reducing agent, i.e., a reagent that lowers the melting point of DNA. Suitable melting point reducing agents are those agents that interfere with the hydrogen bonding interaction of two nucleotides, where representative base pair destabilization agents include: betaine, formamide, urea, thiourea, acetamide, methylurea, glycinamide, and the like, where betaine is a preferred agent. The melting point reducing agent will typically be present in amounts ranging from about 20 to 500 mM, usually from about 50 to 200 mM and more usually from about 80 to 150 mM.

In preparing the PCR reaction mixture, the various constituent components may be combined in any convenient order. For example, the buffer may be combined with primer, polymerase and then template DNA, or all of the various constituent components may be combined at the same time to produce the reaction mixture.

Following preparation of the PCR reaction mixture, it is subjected to a plurality of reaction cycles, where each reaction cycle comprises: (1) a denaturation step, (2) an annealing step, and (3) a polymerization step. The number of reaction cycles will vary depending on the application being performed, but will usually be at least 15, more usually at least 20 and may be as high as 60 or higher, where the number of different cycles will typically range from about 20 to 40. For methods where more than about 25, usually more than about 30 cycles are performed, it may be convenient or desirable to introduce additional polymerase into the reaction mixture such that conditions suitable for enzymatic primer extension are maintained.

The denaturation step comprises heating the reaction mixture to an elevated temperature and maintaining the mixture at the elevated temperature for a period of time sufficient for any double stranded or hybridized nucleic acid present in the reaction mixture to dissociate. For denaturation, the temperature of the reaction mixture will usually be raised to, and maintained at, a temperature ranging from about 85 to 100°C usually from about 90 to 98°C, and more usually from about 93 to 96°C for a period of time ranging from about 3 to 120 seconds, usually from about 5 to 30 seconds.

Following denaturation, the PCR reaction mixture will be subjected to conditions sufficient for primer annealing to template DNA present in the mixture. The temperature to which the reaction mixture is lowered to achieve these conditions will usually be chosen to provide optimal efficiency and specificity, and will generally range from about 50 to 75°C, usually from about 55 to 70°C and more usually from about 60 to 68°C. Annealing conditions will be maintained for a period of time ranging from about 15 seconds to 30 minutes, usually from about 30 seconds to 5 minutes.

Following annealing of primer to template DNA or during annealing of primer to template DNA, the reaction mixture will be subjected to conditions sufficient to provide for polymerization of nucleotides to the primer ends in manner such that the primer is extended in a 5' to 3' direction using the DNA to which it is hybridized as a template, i.e. conditions sufficient for enzymatic production of primer extension product. To achieve polymerization conditions, the temperature of the reaction mixture will typically be raised to or maintained at a temperature ranging from about 65 to 75°C, usually from about 67 to 73°C and maintained for a period of time ranging from about 15 seconds to 20 minutes, usually from about 30 seconds to 5 minutes.

The above cycles of denaturation, annealing and polymerization may be performed using an automated device, typically known as a thermal cycler. Thermal cyclers that may be employed are described in U.S. Patent Nos. 5,612,473; 5,602,756; 5,538,871; and 5,475,610.

Based on all the previous criteria, a series of primers were produced and validated by PCR using genomic DNA from normal individuals. Knowledge of suitable primers will necessarily define the corresponding PCR-produced probes in accordance with the invention. Thus, adjacent pairs of sequences identified as SEQ ID Nos. 429-446 and 480-613, beginning with SEQ ID No. 429, are respective forward/reverse PCR primers developed for the production of specific useful probes. Hence, a useful probe may be produced using a combination of SEQ ID Nos. 429 and 430, and additional probes are defined by the succeeding pairs of adjacent SEQ IDs. Broadly speaking, certain preferred probes of the invention should have at least about 80% sequence identity, and more preferably about 90% sequence identity, relative to the probes defined by the above-described adjacent pairs of primer sequences.

In addition to the PCR, DNA fragments corresponding to unique sequences can also be obtained by a variety of other methods, including but not limited to deletion mutagenesis, restriction digestion, direct synthesis and DNA ligation.

If the genomic fragment is obtained by amplification or purification from DNA containing repetitive sequences, the fragment must then be purified prior to labeling and hybridization. Purification of homogeneously-sized DNA fragments can be accomplished by a variety of methods, including but not limited to electrophoresis and high pressure liquid chromatography. In the preferred method, amplified fragments are separated according to size by gel electrophoreses in Seakem LE Agarose using Tris Acetate buffer (Sambrook, Fritsch & Maniatis, Molecular Cloning: A Laboratory Manual [Cold Spring Harbor Laboratory Press, 1989]), stained with a dye such as ethidium bromide or Syber-Green, visualized with ultraviolet light (300 nm), excised from the gel using a scalpel. Each DNA fragment is then recovered from the gel fragment using a Micro-con 100 (Millipore, Watertown, MA) column by spin centrifugation.

Phenol-chloroform extraction of the amplified DNA is not an adequate method of purification. When this approach was tested, this purification technique resulted in nonspecific hybridization to all chromosomes along their entire length, which is consistent with the pattern produced by hybridization of repetitive sequences (data not shown). This occurs because, during the PCR process, DNA polymerase extends the replicated strand past the position of the second primer into adjacent repetitive sequences if the initial template contains genomic DNA sequences. These extension products which are longer than the amplification product, are present in all such PCR reactions. Since, in the present method, repetitive sequences are adjacent to the segments being amplified, the extension products are likely to contain such sequences. Phenol-chloroform extraction of PCR reactions does not remove such extension products. PCR reaction mixtures containing these sequences may hybridize to repetitive genomic DNA in addition to the target sequence. Hence, isolation of the purified genomic amplification fragment (whether it is obtained directly from genomic DNA or by PCR), is a preferred embodiment of the subject invention and would not be obvious to one skilled in the art.

Insertion of the purified fragments into plasmids, bacteriophages, or artificial chromosome cloning vehicles capable of being propagated in *E. coli*, yeast, or other species

may be desirable to reduce the cost and labor required for repeated preparation of single copy DNA probes. A variety of cloning vectors have been optimized for rapid ligation and selection for vectors containing PCR products (for example: U.S. Patent Nos. 5,487,993 and 5,766,891). If the probe will be used in multiple hybridizations, then the cloned recombinant form will be less expensive to produce in large quantities than by iterative PCR amplification from the same genomic DNA template. In addition, genomic insert in the cloned probe does not have to be isolated during purification, since the fragment recombined with vector is propagated in the absence of any other genomic DNA that could potentially contain repetitive sequences. Finally, the cloned vehicle provides a potentially inexhaustible source of probe, whereas natural genomic DNA templates may have to be reisolated from cell lines or from other sources. Single copy DNA fragments obtained by PCR amplification as described above are isolated according to size by gel electrophoresis and purified by columns as is well known in the art.

These fragments are then labeled with nonisotopic identifying label such as a fluorophore, an enzymatic conjugate, or one selected from the group consisting of biotin or other moieties recognized by avidin, streptavidin, or specific antibodies. There are several types of non-isotopic identifying labels. One type is a label which is chemically bound to the probe and serves as the means for identification and localization directly. An example of this type would be a fluorochrome moiety which upon application of radiation of proper wavelengths will become excited into a high energy state and emit fluorescent light. The probes can be synthesized chemically or preferably be prepared using the methods of nick-translation (Rigby et al., *J. Mol. Biol.*, **113:**237-251, (1977)) or Klenow labeling (Feinberg et al., *Anal. Biochem.*, **137:**266-267, (1984)) in the conventional manner using a reactant comprising the identifying label of choice (but not limited to) conjugated to a nucleotide such as dATP or dUTP. The fragments are either directly labeled with a fluorophore-tagged nucleotide or indirectly labeled by binding the labeled duplex to a fluorescently-labeled antibody that recognizes the modified nucleotide that is incorporated into the fragment as described below. Nick-translations (100 µl reaction) utilize endonuclease-free DNA polymerase I (Roche Molecular Biochemicals, Indianapolis, IN) and DNase I (Worthington Biochemical Corporation, Lakewood, NJ). Each fragment is combined with DNA polymerase (20 units/microgram DNA), DNase I (10 microgram/100 µl reaction), labeled

nucleotide (0.05 mm final) and nick translation buffer. The reaction is performed at 15°C for 45 minutes to 2 hours and yields a variety of labeled probe fragments of different nucleotide sizes in the 100 to 500 bp size range.

Other methods for labeling and detecting probes in common use may be applied to the single-copy DNA probes produced by the present method. These include: fluorochrome labels (which resolve labeling on individual chromatids which serves as an affirmation that hybridization occurred unequivocably, and further allows detection precisely at site of hybridization rather than at some distance away), chemical reagents which yields an identifiable change when combined with the proper reactants (for example, alkaline phosphatase, horseradish peroxidase and galactosidase, each of which reacts and provide a detectable color change that identifies the presence and position of the target sequence), and indirect linkage mechanism of specifically binding entities (such as the biotin-avidin system in which the probe is preferably joined to biotin by conventional methods and added to an avidin- or streptavidin-conjugated fluorochrome or enzyme which provides the specificity for attaching the fluorochrome or enzyme to the probe).

It will be recognized that other identifying labels may also be used with the described probes. These include fluorescent compositions such as energy transfer groups, conjugated proteins, antibodies and antigens, or radioactive isotopes.

Chromosomal hybridization and detection are a preferred use of DNA probes generated by the present invention. DNA probes generated by the present method may be hybridized either directly to complementary nucleic acids in cells (in situ hybridization) or to nucleic acids immobilized on a substrate. A preferred use of the method is *in situ* hybridization, which is well known in the art, being described in U.S. Patent. Nos. 5,985,549; 5,447,841; 5,756,696; 5,869,237. Based on early work of Gall and Pardue (*Proc. Natl. Acad. Sci.,* **63**:378-383, 1969), isotopic *in situ* hybridization was established in the 1970s (see Gerhard et al., *Proc. Natl. Acad. Sci.,* **78**:3755-3759, 1981 and Harper et al., *Proc. Natl. Acad. Sci.,* **78**:4458-4460, 1981 as examples) and subsequently nonisotopic *in situ* hybridization was established. The technique of nonisotopic in situ hybridization is reviewed and a protocol is provided for use in chromosomal hybridization by Knoll and Lichter, in *Current Protocols in Human Genetics*, Vol. 1, Unit 4.3 (Green-Wiley, New York, 1994) and in U.S. Patent No. 5,985,549. The technique relies on the formation of duplex

nucleic acid species, in which one strand is derived from a labeled probe molecule and the second strand comprises the target to be detected. Target molecules may comprise chromosomes or cellular nucleic acids. Numerous methods have been developed to label the probe and visualize the duplex.

The method of the present invention is intended to be used with any nucleic acid target containing repetitive sequences. The sample containing the target nucleic acid sequence can be prepared from cellular nuclei, morphologically intact cells (or tissues), chromosomes, other cellular material components, or synthetically produced nucleic acids. The samples may be obtained from the fluids or tissues of a mammal, preferably human, which are suspected of being afflicted with a disease or disorder either from a biopsy or post-mortem, or from plants.

As an example, chromosomal preparations can be made in the following manner: phytohemagglutinin-stimulated peripheral lymphocytes are cultured in RPMI 1640 medium containing 10% fetal calf serum for 72 hours at 37°C. Ethidium bromide (100ug/10ml final) is added 1-1/2 hours prior to harvest. Colcemid (1ug/10ml final) is added during the final 20 min of incubation with ethidium bromide. The cells are then pelleted by centrifugation and incubated preferably in 0.075 M KCl at 37°C for about 20 minutes. Cells are then pelleted again and fixed in 3 changes of Carnoy's fixative (3:1 methanol: acetic acid volumetric ratio) using conventional cytogenetic techniques. [For a review of chromosome preparation from peripheral blood cells, see Bangs and Donlon in Dracopoli et al., eds., *Current Protocols in Human Genetics*, Vol. 1, Unit 4.1 (Green-Wiley, New York, 1994)]. The nuclei or cells in suspension can then be dropped onto clear glass coverslips or microscope slides in a humid environment to promote chromosome spreading. The coverslips or microscope slides can then be preferentially air dried overnight, aged or stored until required for use in *in situ* hybridization.

Immediately prior to chromosomal denaturation in the *in situ* hybridization procedure, the dried or stored chromosome preparations can be pretreated in prewarmed 2 x SSC (components are in Sambrook, Fritsch & Maniatis, Molecular Cloning: A Laboratory Manual [Cold Spring Harbor Laboratory Press, 1989]) for 30 minutes at 37°C followed by dehydration in an ethanol series (2 minutes each in 70%, 80%, 90% and 100% ethanol).

When the target nucleic acid sequence is DNA, DNA in the sample can be denatured by heat or alkali. [See Harper et al., *Proc. Natl. Acad. Sci.,* **78**:4458-60 (1981), for alkali denaturation and Gall et al., *Proc. Natl. Acad. Sci.,* **63**:370-383 (1969)]. Denaturation is carried out so that the DNA strands are separated with minimal shearing, degradation or oxidation.

In the preferred current method, the labeled single copy probe is resuspended in deionized formamide and denatured at 70-75°C. The chromosomal template is denatured in a solution containing 70% formamide/2 x SSC, pH 7.0 followed by dehydration in an ethanol series (2 minutes each in cold 70% ethanol and room temperature 80, 90 and 100% ethanol). Hybridization of the labeled probe to the corresponding template is carried out in a solution containing 50% formamide/2 x SSC/10% dextran sulfate/BSA [bovine serum albumin; 1 mg/ml final] for a few hours to overnight. The length of time depending on the complexity of the probe that is utilized. After hybridization, non-hybridizing excess probe is removed by a washing procedure. The duplexes are treated with a series of 15-30 minute washes: first with a solution of 50% formamide/2xSSC at 39-45°C, then 2 x SSC at 39-45°C, followed by a 15-30 minute wash at room temperature in 1 x SSC. The hybridized sequences are detected by relevant means. For example, digoxigenin-dUTP can be but is not limited to detection by an antibody to digoxigenin such as rhodamine or fluorescein conjugated antibody (Roche Molecular Biochemicals, Indianapolis, IN). Following detection, spurious detection reagents are removed by washing in varying SSC and SSC/triton-X concentrations, the chromosomes are counterstained with a dye such as DAPI and the hybridized preparation is mounted in an antifade solution such as Vectashield (Vector Laboratories, Burlingame, CA). The cells are examined by fluorescence microscopy with the appropriate filter sets and imaged with a charge coupled device (CCD).

An important aspect of the present invention is that the probe or target DNA does not require pre-reaction with a non-specific nucleic acid competitor such as purified repetitive DNA or that the probe does not require experimental verification that the single copy fragments or recombinant cloned probes do not contain repetitive sequences (U.S. Patent Nos. 5,985,549; 5,447,841; 5,663,319; 5,756,696) because the probes are single copies without repetitive elements. This results in a significantly improved signal to noise ratio. A signal to noise ratio is defined as a ratio of the probability of the probe detecting a *bona*

fide signal of hybridization of the target nucleic acid sequence to that of the probability of detecting the background caused by non-specific binding of the labeled probe.

The hybridization reactions carried out using the probes of the invention are themselves essentially conventional. As indicated, two exemplary types of hybridizations are the Southern blot and FISH techniques, well known to those skilled in the art. However, the visual patterns resulting from use of the probes, termed indicator patterns, are extremely useful tools for cytogenetic analyses, especially molecular cytogenetic analyses. These indicator patterns facilitate microscopic and/or flow cytometric identification of normal and abnormal chromosomes and characterization of the genetic abnormalities. Since multiple compatible methods of probe visualization are available, the binding patterns of different components of the probes can be distinguished, for example, by color. Thus, the invention is capable of producing virtually any desired indicator pattern on the chromosomes visualized with one or more colors (a multi-color indicator pattern) and/or other indicator methods.

Preferred indicator patterns derived from using the probes of the invention comprise one or more "bands," meaning a reference point in a genome comprising a target DNA sequence with a probe bound thereto, and wherein the resulting duplex is detectable by some indicator. Depending on hybridization washes and the detection conditions, a band can extend from the narrow context of a sequence providing a reliable signal to a single chromosome region to multiple regions on single or plural chromosomes. The indicator bands from the probes hereof are to be distinguished from bands produced by pretreatment and chemical staining. The probe-produced bands of the present invention are based upon the complementarity of the DNA sequences, whereas bands produced by chemical staining depend upon natural characteristics of the chromosomes (such as structure or protein composition), but not by hybridization to the DNA sequences thereof. Furthermore, chemical staining techniques are useful only in connection with metaphase chromosomes, whereas the probe-produced bands of the present invention are useful for both metaphase and interphase chromosomes.

The following examples set forth the preferred techniques employed for the development, generation, labeling and use of specific DNA probes designed to hybridize to a target DNA sequence in a genome. It is to be understood, however, that these examples

are provided by way of illustration and nothing therein should be taken as a limitation upon the overall scope of the invention.

Example 1

Development of HIRA Gene Probe

A known genetic disorder on human chromosome 22 involves a deletion of one HIRA gene in chromosome band 22q11.2, i.e., in normal individuals; there are two copies of the HIRA gene, whereas in affected individuals, only one copy is present. This deletion is considered to be a cause of haploinsufficiency syndromes such as DiGeorge and Velo-Cardio-Facial Syndromes (VCFS), because insufficient amounts of gene product(s) may disrupt normal embryonic development (Fibison et al., *Amer. J. Hum. Genet.*, **46:**888-95 (1990); Consevage et al., *Amer. J. Cardiol.*, **77:**1023-1205 (1996)). Other syndromes including Cat Eye Syndrome and derivative chromosome 22 syndrome result from an excess of genomic sequences from this region (Mears et al., *Amer. J. Hum. Genet.*, **55:**134-142 (1994); Knoll et al., *Amer. J. Med. Genet.*, **55:**221-224 (1995)). Typically individuals with these syndromes have supernumerary derivative chromosome 22s.

Initially, a computer-based search using the search term "HIRA" was performed using Entrez Nucleotide software at the National Library of Medicine website. This identified a series of cDNA sequences for the HIRA gene in GenBank. The full length cDNA sequence was selected (GenBank Accession No. X81844), having 3859 bp. This cDNA sequence was then compared with the genome sequence which included draft sequences at the National Library of Medicine (http://www.ncbi.nlm.nih.gov/genome/seq/page.cgi?F= HsBlast.html&&ORG=Hs). This was done in order to determine whether genomic sequences of sufficient length were available for probe development. This comparison confirmed that the entire HIRA genomic sequence was known, and that the coding sequence interval spanned a length of 100,836 bp in the chromosome. Since the available contiguous genomic sequence in GenBank exceeded the length of the coding interval, it was possible to select an interval longer than the coding region in order to include sequences from the gene promoter at the 5' end and untranslated sequences and polyadenylation signal at the 3' end. A total genomic interval of approximately 103 kb was thus

selected. Position 1 of this ~103 kb interval corresponds to position 798,334 in GenBank Accession number NT_001039.

In the next step, the selected 103 kb genomic interval was compared with known high-complexity repeat sequence family members or consensus sequences that are aligned with the test genomic sequences (SEQ ID Nos. 1-428) and all combinations of low-complexity tandem repeat sequences of at least 17 nucleotides in length (mono-, di-, tri-, and tetranucleotide units) known to be present in the human genome (SEQ ID Nos. 447-479). This comparison was done using the publicly available CENSOR program which can be found at the Genetic Information Research Institute website, www.girinst.org. This program utilizes the Smith-Waterman global alignment comparison algorithm to determine the locations and distribution of repeat sequences within the genomic interval. A Smith-Waterman alignment of repetitive with genomic sequences was performed with the following parameters: Length of margin sequence: 50 nt, minimum length to extract insertion: 12 nt, minimum margin to combine matching fragments: 30, similarity threshold: 22, similarity threshold to always keep match: 35, ratio threshold: 2.8, relative similarity threshold: 2.8, gap constant D1: 2.95, gap constant D2: 1.90, and mismatch penalty: -1.0. This analysis generated the following table, which details the coordinates of repetitive sequence family members found in and adjacent to the human HIRA gene coding sequence.

Table 1

HIRA position (bp)				Position (bp) in Seq.Listing Corresponding to HIRA Match	
Begin	End	Repeat Family	SEQ ID NO.	Begin	End
798411	798434	(AC)	452	1	24
798983	799395	MLT2A1	444	1	434
801257	801348	CHESHIRE_A	420	132	223
801367	801729	L1ME_ORF2	425	757	1089
801746	802032	Alu-Jb	2	2	289
802033	802308	L1ME_ORF2	425	1090	1380
802355	802434	L1MB6_5	77	1629	1710
802448	802798	L1M3D_5	66	996	1348
802811	803100	Alu-Y	2	1	290
803104	803189	L1M3D_5	66	907	995
803199	803454	Alu-Jb	2	5	290
803472	803545	Alu-Spqxz	2	2	76
803548	804061	L1MEC_5	345	1860	2392
804079	804365	Alu-Sz	2	6	290
804476	804559	L1P_MA2	348	6242	6321
804625	804885	L1ME_ORF2	425	2287	2568
804936	804997	MLT1E2	106	198	260
805011	805077	MLT1E1	105	420	484

Begin	End	Repeat Family	SEQ ID NO.	Begin	End
805110	805211	L1PBA_5	359	103	204
805212	805862	L1PBA_5	359	1089	1738
805933	805989	Alu-J	2	234	290
805991	806489	L1PBA_5	359	1749	2247
806510	806624	L1	59	1659	1773
806628	806917	Alu-Sz	2	1	290
806919	807254	L1M2_5	61	2377	2716
807301	808176	L1P_MA2	348	3516	4425
808179	808469	Alu-Sz	2	1	290
808476	808734	L1	59	3268	3525
808735	809426	L1ME_ORF2	425	1411	2105
809429	809860	L1P_MA2	348	5607	6044
809861	809993	Alu-Jb	2	2	134
809996	810282	Alu-Jb	2	2	290
810345	811040	L1	59	4711	5402
811041	811221	L1PB3	358	151	333
811226	811513	Alu-Sx	2	1	287
811515	812032	L1PB1	357	330	863
812096	812394	Alu-Jb	2	1	288
812474	812698	Alu-Jb	2	5	229
812721	812836	Alu-Jo	2	2	117
812862	812901	L1P_MA2	348	4315	4354
812903	813078	L1	59	3028	3222
813079	814102	L1ME_ORF2	425	1113	2166
814323	814410	MER1B	315	242	337
814411	814557	CHARLIE3	7	1	281
814780	814916	L1MB7	78	9	143
815061	815181	Alu-Y	2	1	134
815420	815452	LTR67	279	99	131
816487	816772	Alu-Sx	2	5	290
817180	817270	L1MCC_5	335	1384	1473
817332	817620	Alu-Sg	2	1	290
817634	817909	Alu-Sq	2	1	288
817943	818227	Alu-Sx	2	2	289
818368	818578	HAL1	18	1346	1547
818631	818791	LINE2	362	2280	2464
818824	818889	Alu-S	2	223	290
818890	819185	LINE2	362	2465	2749
819328	819450	LINE2	362	1925	2049
819565	819757	LINE2	362	2273	2498
823604	823892	Alu-Jo	2	2	290
826836	827042	Alu-Sxzg	2	84	290
827922	827977	MIR	99	105	160
830762	831371	L1MEC_5	345	1498	2123
831396	831685	Alu-Sx	2	2	290
831687	831774	L1MEC_5	345	2117	2205
831778	832066	Alu-Sx	2	1	290
832155	832288	Alu-FLA	2	5	134
832317	832431	L1MC2	79	666	786
832442	832735	Alu-Sz	2	1	289
832742	832992	L1MC2	79	787	1077
833004	833170	L1ME_ORF2	425	172	340
833177	834590	TIGGER1	148	1	1477
834592	834642	Alu-Jb	2	156	207
834799	834877	Alu-Jb	2	208	290
834907	835194	Alu-Y	2	1	289
835198	835590	TIGGER1	148	1468	1900
835597	835888	Alu-Sx	2	1	290
835946	835979	L1P_MA2	348	4654	4689
836060	836177	MER2	316	229	345
836203	836486	Alu-Sx	2	7	290
836497	836712	MER2	316	1	228
838478	838760	Alu-Sz	2	1	288
838822	839069	Alu-Sx	2	1	288
839086	839373	Alu-Sz	2	1	289
840297	840926	L1MB7	78	269	915
841062	841306	L1MB7	78	7	249
841323	841382	L1ME_ORF2	425	3053	3116

Begin	End	Repeat Family	SEQ ID NO.	Begin	End
841408	841697	Alu-Sq	2	1	290
841705	841828	Alu-Jo	2	1	136
841829	842012	L1ME_ORF2	425	2870	3052
842744	842871	MER86	239	51	183
842879	843107	Alu-Spqxz	2	3	230
843109	843271	Alu-Jo	2	9	175
847056	847210	MER104	293	1	179
847256	847351	L1ME4	343	128	224
847413	847551	MIR	99	65	218
847570	847695	L1ME4	343	1	127
847865	848137	Alu-Y	2	1	290
848171	848458	Alu-Sg	2	1	290
848493	848564	L1PA7	355	35	105
848646	848928	Alu-Sc	2	5	290
849186	849435	L1ME_ORF2	425	2527	2796
849450	849745	Alu-Sx	2	5	289
850114	850249	LIP_MA2	348	5447	5610
850250	850761	L1	59	3478	4017
850824	850942	L1ME_ORF2	425	1128	1265
851588	851614	(T)	449 (complement)	1	27
851749	851881	L1ME2	341	357	523
852607	852853	L1MA10	72	664	918
852863	853156	Alu-Sc	2	1	290
853176	853211	L1MA10	72	628	663
853212	853267	L1MA9	75	987	1041
853491	853779	Alu-Sz	2	1	290
859137	859435	Alu-Sx	2	1	290
859436	859456	(A)	449	1	21
859570	859805	L1ME3A	342	215	442
859806	860289	L1ME2	341	375	879
860318	860605	Alu-Y	2	1	290
862194	862481	Alu-Sg	2	1	290
865060	865350	Alu-Sq	2	1	290
867521	867800	Alu-Jb	2	1	288
867836	867876	MIR	99	157	196
869546	869802	LINE2	362	123	413
869923	870118	LINE2	362	1251	1450
870124	870202	Alu-J	2	48	132
870203	870296	LINE2	362	1451	1592
870316	870666	LINE2	362	1708	2097
871000	871075	LINE2	362	2617	2736
871650	871935	Alu-Jo	2	1	290
871936	871960	(GAAAAA)		4	28
872154	872444	Alu-Sc	2	1	289
874867	874990	L1MB7	78	529	676
878120	878408	Alu-Sx	2	1	290
881003	881054	MLT1G	109	217	268
881130	881266	MLT1G	109	269	480
881293	881346	MLT1G	109	415	469
881762	881891	LINE2B	363	85	229
882448	882740	Alu-Sb0	2	1	290
883566	883716	Alu-Sz	2	1	288
883782	883977	Alu-Sc	2	2	290
883988	884329	LIP_MA2	348	5600	5935
884333	884623	Alu-Sp	2	1	290
884624	885134	L1ME_ORF2	425	2431	2975
885160	885456	Alu-Jb	2	9	290
885460	885742	L1ME_ORF2	425	2949	3252
885744	886031	Alu-Sx	2	1	288
886032	886082	Alu-Sp	2	291	341
886083	886166	L1MB7	78	137	220
886168	886454	Alu-Sc	2	1	290
886535	887059	L1MB7	78	345	901
887169	887460	Alu-Y	2	1	289
887485	887748	L1MD2	337	794	1072
887752	887779	LOR1I	366	395	422
888253	888318	LINE2	362	2440	2505
888385	888548	LINE2	362	2579	2739

Begin	End	Repeat Family	SEQ ID NO.	Begin	End
888865	888893	LOR1I	366	394	422
889006	889296	Alu-Jb	2	5	290
889446	889548	Alu-Jo	2	188	290
889549	889677	L1PB3	358	770	897
889842	890133	Alu-Sq	2	1	290
890515	890797	Alu-Sz	2	1	283
890858	890972	L1ME2	341	769	885
890986	891024	LOR1I	366	396	434
891028	891063	LTR66	266	173	207
891126	891536	LINE2	362	1980	2452
891545	891670	LTR16A1	382	9	128
891688	891963	LTR16A	381	146	429
892907	893013	LINE2	362	2636	2747
893851	893924	MLT1L	119	47	119
894528	894849	Alu-Sx	2	1	290
895825	895903	LINE2	362	2592	2664
895912	896083	MER20	317	46	216
897067	897299	MER20	317	2	217
897492	897624	Alu-FLA	2	2	136
897977	898261	Alu-Sc	2	1	290

The lengths of the non-repetitive intervals were calculated from these data. For example, a non-repetitive interval of 5358 bp was determined between coordinate positions 853779 and 859137 which delineate the boundaries of adjacent Alu-Sz and Alu-Sx repetitive elements. Next, the non-repetitive intervals were sorted based on their respective lengths. Four of these non-repetitive intervals were selected for probe development, namely the above-referenced 5358 bp sequence, a 3847 bp sequence (coordinates 819757 and 823604), a 3785 bp sequence (coordinates 843271 and 847056), and a 3130 bp sequence (coordinates 874990 and 878120).

In the next step, the long PCR technique was used to amplify portions of the four identified single copy intervals. The technique followed for amplification of the 5358 bp interval is described in detail below. Similar techniques were followed for amplification of the remaining three single copy intervals.

Probes of maximal length were desired for FISH experiments. However, in order to optimize the PCR reaction that generated these probes, other constraints had to be met, which resulted in amplification products somewhat shorter than the entire non-repetitive sequence interval. The Prime computer program was employed to optimize the selection of primers for PCR (Genetics Computer Group software package, Madison WI). The PCR primers which were optimized for long PCR were constrained as follows: size of 30-35 nucleotides; GC content of 50-80%; melting temperature of 65-70°C; the primer was not permitted to self-anneal at the 3' end with hairpins of greater than 8 nucleotides; the primer was not permitted to self-anneal at any position with greater than 14; and the primer was

permitted to anneal only at a single position in the target sequence and primer-primer annealing was limited at the 3' end to less than 8 bp and at any other point less than 14 bp. In addition, certain constraints were applied to the amplified PCR product in order to optimize long PCR: length of 5100-5358 nucleotides; GC content of 40-60%; melting temperature of 70-95°C; difference in forward and reverse primer melting points less than 2°C. This yielded a possible 517 forward primers and 382 reverse primers, and a total of 928 possible products. The Prime program using the foregoing constraints rank ordered potential primer pairs. The top ranked primers were selected for synthesis, as set forth in the following Table 2. These primers were commercially produced (Oligos, Etc., Wilsonville, OR).

Table 2

Gene	Chromosome Band	GenBank Accession No., Chromosome Genomic Sequence	Coordinates of Longest Single Copy Intervals, Beginning/End	Forward PCR Primer Coordinates, Beginning/End	Reverse PCR Primer Coordinates, Beginning/End	PCR Primer SEQ ID Nos., Forward/Reverse	Probe Length (bp)
HIRA	22q11.2	NT_001039	853779/859137	853946/853975	859116/859085	429/430	5170
HIRA	22q11.2	NT_001039	819757/823604	819901/819933	823592/823559	431/432	3691
HIRA	22q11.2	NT_001039	843271/847056	843602/843631	846946/846915	433/434	3344
HIRA	22q11.2	NT_001039	874990/878120	875226/875257	878074/878042	435/436	2848

Using these primers, a long PCR reaction (50 µl) was performed using 1 microgram of high molecular weight genomic DNA (purified by phenol extraction) and 200 µM of each oligonucleotide primer to amplify the 5170 bp probe. Specifically, high fidelity DNA polymerase (LA-Taq, Takara Chemical Co.) was employed using the following thermal cycling protocol:

Step 1 - 94°C - 5 minutes
Step 2 - 98°C - 20 seconds
Step 3 - 65°C - 7 minutes
Step 4 - 14 times to Step 2
Step 5 - 98°C - 20 seconds
Step 6 - 65°C - 7 minutes + 15 s/cycle
Step 7 - 14 times to Step 5
Step 8 - 72°C - 10 minutes
Step 9 - 0°C
Step 10 - END

Because amplification of the 5170 bp probe is less efficient than amplification of shorter fragments, the initial PCR reaction did not yield sufficient quantities of probe for multiple hybridization experiments. Therefore, a 4 µl aliquot of the original DNA amplification reaction was reamplified using the following protocol: Step 1 - 94°C - 1.5 minutes, followed by Steps 2-10 of the original PCR reaction. Sufficient quantities of the 5170 bp probe were obtained. An alternative to reamplification is to increase Step 7 by at least 10 cycles.

The amplified product was then purified by gel electrophoresis followed by column chromatography. First, the amplified product was separated on a 0.8% Seakem LE agarose gel (FC Bioproducts) in 1X modified TAE buffer. The gel was then stained with ethidium bromide and visualized with UV light. The fragment corresponding to the correct interval size was excised in an Ultrafree-DA spin column (Millipore) and centrifuged at 5000g for 10 minutes. The DNA was recovered in solution and precipitated in 1/10 V NaOAc and 2.5 V 95% EtOH (overnight) at -20°C. The precipitated DNA was then centrifuged, rinsed with cold 70% EtOH, air dried and

resuspended in 20 µl of sterile deionized water. The DNA was checked on a 0.8% agarose gel (Sigma) to determine DNA concentration.

The detailed probe labeling, hybridization, removal of non-specifically bound probe, and probe detection procedures are described by Knoll and Lichter, In: Dracopoli et al., (eds), "Current Protocols in Human Genetics Volume 1", Unit 4.3 (Green-Wiley, New York,1994). Briefly, in order to label the probe, a standard nick translation reaction was carried out (Rigby et al., *J. Mol. Biol.*, **113**:237-251, (1977)) using digoxigenin-11-dUTP as the label. This yielded a series of overlapping 300-500 bp labeled fragments, which together comprised the 5170 bp probe.

The labeled probe fragments were then precipitated by adding 1/10 V NaOAc plus 2.5 V 95% EtOH and carrier DNA (overnight, -20°C). On the following day, the precipitated DNA was centrifuged, lyophilized, and resuspended in deionized sterile water at a concentration of 125 ng/20 µl.

A comparison set of hybridizations were carried out with normal denatured human metaphase chromosomes, using the labeled probe fragments with and without blocking nucleic acid of the type described in U.S. Patents Nos. 5,447,841, 5,663,319 and 5,756,696. Twenty µl of resuspended labeled probe was then lyophilized and resuspended in 10 µl of deionized formamide and denatured for 5 minutes at 70-75°C to yield single-stranded nucleic acids. For comparison, probes were pre-reacted with purified repetitive DNA by adding 125 ng (or 20 µl) of labeled probe to 10 micrograms of C_0t 1 DNA (Life Technologies) and lyophilizing the mixture. This mixture was then denatured for 5 minutes at 70°C followed by pre-reaction (or pre-annealing) for 30 minutes at 37°C to convert the single stranded repetitive sequences in the probe to double stranded nucleic acid. This disables the hybridization between the sequences and the chromosome as target DNA template.

Subsequently, the denatured probes with or without purified repetitive DNA (i.e., C_0t 1) were mixed with 1 V prewarmed hybridization solution (comprised of 4 x SSC/2 mg/ml nuclease free bovine serum albumin/20% dextran sulfate/30% sterile deionized water) and overlaid onto denatured target DNA. The chromosomal target DNA, fixed to a microscope slide had been denatured at 72°C for 2 minutes in 50% formamide/2 x SSC. A coverslip was placed over the probe hybridization mixture on

the slide, sealed with nail polish enamel to prevent evaporation and placed in a moist chamber at 39°C overnight.

Following hybridization, non-specifically bound probe was washed off with varying stringencies of salt concentration and temperature. The labeled probes, pre-reacted to disable repetitive sequence hybridization, and the probes without such pre-reaction were detected with rhodamine-labeled antibody to digoxigenin-11-dUTP, using a conventional FISH protocol (Knoll and Lichter, *Current Protocol in Human Genetics*, Vol. 1, Unit 4.3, Green-Wiley, New York, 1994). Chromosomal DNA was counter-stained with DAPI. The cell preparations on microscope slides were then mounted in antifade solution (such as Vectashield, Vector Laboratories, Burlingame, CA) and visually examined using a fluorescence microscope with the appropriate fluorochrome filter sets. Figs. 1 and 2 are photographs illustrating the results of the comparative hybridizations, where Fig. 1 is the hybridization with the blocking repetitive sequences, while Fig. 2 is the hybridization without pre-reaction with purified repetitive DNA. These photographs depict hybridization to both HIRA alleles on two normal chromosome 22q11.2 regions. A comparison of the photographs demonstrates that the presence of the blocking repetitive sequences is unnecessary using the probes of the present invention.

The remaining three probes identified in Table 2 were PCR-amplified and labeled as described above. These probes were used in a series of FISH experiments to determine the efficacy of the probes. Thus, all four probes were used together without pre-annealing of potentially repetitive sequences (Fig. 6), and a combination of the three shortest probes were used on cells from a patient affected with DiGeorge/VCFS with a previously confirmed deletion (Fig. 12). In the Fig. 6 photograph, the probe was hybridized to a single region of both chromosome 22s in a normal individual (arrows) In Fig. 12, only one chromosome 22 hybridized (arrow). The other chromosome 22, as indicated by a star, has a deletion of this region and does not hybridize to the probe.

Example 2

Development of NECDIN and CDC2L1 Gene Probes

The techniques described in Example 1 were used to develop a series of probes for detecting known genetic disorders on chromosome 1 (Monosomy 1p36.3 syndrome; Slavotinek et al., *J. Med. Genet.*, **36:**657-63 (1999)) and on chromosome 15 (Prader-Willi and Angelman Syndromes). Approximately 70% of patients with Prader-Willi or Angelman syndrome exhibit hemizygous deletions of the sequence containing the NECDIN gene (Knoll et al., *Amer. J. Med. Genet.*, **32:**285-290 (1989); Nicholls et al., *Amer. J. Med. Genet.*, **33:**66-77 (1989)). The presence of excess copies of this gene is diagnostic for an abnormal phenotype in patients with interstitial duplication or a supernumerary derivative or dicentric chromosome 15 (Cheng et al., *Amer. J. Hum. Genet.*, **55:**753-759, (1994); Repetto et al., *Am. J. Med. Genet.*, **79:**82-89, (1998)). The following Table 3 sets forth the deduced single copy intervals, PCR primer coordinates, SEQ ID Nos., and the lengths of the resultant probes.

Table 3

Gene	Chromosome Band	GenBank Accession No., Chromosome Genomic Sequence	Coordinates of Longest Single Copy Intervals, Beginning/End	Forward PCR Primer Coordinates, Beginning/End	Reverse PCR Primer Coordinates, Beginning/End	PCR Primer SEQ ID Nos., Forward/Reverse	Probe Length (bp)
CDC2L1[1]	1p36.3	AL031282	8823/17757	9137/9167	13960/13931	444/443	4823
CDC2L1[1]	1p36.3	AL031282	8823/17757	13028/13057	17752/17720	445/446	4724
NECDIN	15q11-q13	AC006596	94498/99152	94501/94535	98567/98601	439/440	4166
NECDIN	15q11-q13	AC006596	68031/75948	72122/72156	75666/75637	437/438	3544
NECDIN	15q11-q13	AC006596	76249/79221	76608/76639	78898/78867	441/442	2290

[1]Two sets of primers were used to generate two DNA probe fragments which, together, spanned the entire interval.

PCR-amplification was performed using the CDC2L1 primers in Table 3, and products were labeled, hybridized and detected as set forth in Example 1. The labeled probes were used in a series of FISH experiments, with images of the hybridizations provided as Figs. 7-10. In the experiment shown in Fig. 7, the longest 4823 bp probe was employed and potential hybridization repetitive sequences was disabled by pre-annealing with purified repetitive DNA. As a comparison, the same probe was used without pre-annealing of purified repetitive DNA (Fig. 8). The hybridizations appear identical demonstrating that the presence of purified repetitive DNA to block repetitive sequence hybridization is unnecessary. In both instances, the chromosomes with one or both chromatids hybridized are indicated by arrows. In the experiments shown in Figs. 9 and 10, the 4823 bp and 4724 bp probes were employed, with (Fig. 9) and without (Fig. 10) pre-annealing of the purified repetitive DNA. Again, pre-reaction of the purified repetitive DNA is shown to be unnecessary using the probes of the invention.

The NECDIN probes were also used in a series of FISH experiments, as shown in Figs. 3-5 and 11. These probes detected DNA sequences between 36 and 62 kb distal of the NECDIN gene. The 3544 bp probe (SEQ ID Nos. 437-438) detected the 3' terminus of the MAGEL2 gene. In Fig. 3, the 3544 bp probe was used on metaphase cells from a normal individual, with pre-annealing using purified repetitive sequences; Fig. 4 is a comparison, without pre-annealing. In Fig. 11, all three probes were used in combination, on metaphase cells from a patient affected with Prader-Willi syndrome known to harbor a deletion of 15q11-q13 sequences on one chromosome 15. The normal homolog is indicated by an arrow and shows hybridization to a single chromatid. The location of the deleted chromosome is indicated by a star. It does not show hybridization with the probe.

The foregoing examples demonstrate that the mixed combinations of DNA fragments give identical hybridization results, as compared with the fragments when used individually. This establishes that none of the fragments used individually or in combination will hybridize to any other location in the genome and hence, are free of repetitive sequences. This provides an additional confirmation of the validity of the present method for the design and production of single copy genomic probes.

Current use of commercial and research genomic probes to detect these disorders requires that hybridization of repetitive sequences be disabled prior to annealing of the probe to metaphase or interphase chromosomes. This increases the number of steps required to perform the protocol and could potentially increase the chances of procedural errors occurring, any of which would be unacceptable in the clinical diagnostic laboratory. The results present in Fig. 7 are comparable to those obtained using related commercially available genomic probes to detect these abnormalities. Hence, these probes will be useful in the detection of these genetic disorders. The probes themselves or in combination with other solutions necessary for hybridization and detections can be provided to clinical laboratories as kits for detection of these genetic disorders.

The probes developed from genomic sequences other than those presented as examples cited herein can also be utilized to detect inherited, sporadic, or acquired chromosomal rearrangements. These rearrangements may correspond to numerous other known genetic abnormalities (including neoplasias) and syndromes besides those examples given above. Hence, the present invention can also be useful for producing probes from genomic regions where no commercial probes are available or the probes are imprecise.

In principle, the present method can be utilized to design, develop and produce single-copy genomic probes for any genomic interval where the DNA sequence is available and where a comprehensive set of repetitive sequence elements in the genome has been cataloged. Such catalogs are currently available for genomes for the following organisms (http://www.girinst.org): *Homo sapiens, Mus musculus, Arabidopsis thaliana, Canorhabditis elegans, Drosophila melanogaster, and Danio rerio*.

Example 3

In this example, a number of probes specific to additional genetic disorders and cytogenetic abnormalities were developed using the principles of the invention. Software was also developed and improved to expedite the process of designing single copy probes (findi.pl, prim_wkg, and prim, referred to above and provided on the

accompanying CD-R). The probes were subsequently tested and their utility confirmed by *in situ* hybridization.

Identification of single copy sequences.

The locations of single copy probe sequences are determined directly from long contiguous genomic DNA sequences. The locations were determined by software that aligns the sequences of repetitive sequence family members with the target genomic sequence. Comparison of the target sequence with previously determined sequences of repetitive family members served to identify and delineate the bounds of repetitive elements within the target. The computer program, RepeatMasker (http://ftp.genome.washing-ton.edu/RM/RepeatMasker.html; Smit A.F.A. & Green P., unpublished results), was used to determine the locations of repetitive sequence families in contiguous genomic sequences, usually ~100 kb in length. RepeatMasker compares a genomic sequence with a compilation of repetitive sequence families present in multiple copies in the human genome (http://www.girinst.org/~server/repbase.html). This repeat sequence database contains representative and consensus sequences for the majority of human repetitive sequence families. The database can be expanded by addition of newly discovered repetitive sequence families (as shown in Example 6).

A Perl script (findi.pl) parsed the coordinates of the boundaries of the repetitive segments from RepeatMasker output, and then deduced and sorted the adjacent single copy intervals by size greater than a parametrized threshold (~2 kb, in most instances). This script determines the locations and lengths of single copy intervals sorted by size from the output file (with the suffix: .out) produced by RepeatMasker, which contains a table of locations and lengths of repeat family elements. The boundaries of adjacent single copy intervals were deduced by subtracting one nucleotide position from the upstream boundary of a repetitive element and adding one nucleotide position to the downstream boundary of the previous element. Single copy intervals with identical upstream and downstream coordinates (1 bp in length) were considered to be adjacent repetitive sequences. Probe sequences were then compared with the human genome sequence database (Altschul et al., *J. Mol. Biol.*, **215**:403-410 (1990)) to determine if there was similarity to sequences elsewhere in the genome (such as duplicons or

triplicons or other less well conserved intervals). Probe sequences that are weakly conserved elsewhere in the genome do not cross-hybridize to those targets.

Oligonucleotide primers were selected for PCR amplification of the longest single copy intervals. A Unix wrapper script (prim_wkg) iteratively modifies the switches in the file containing the command to design primers (prim), thus optimizing primer selection by changing the following parameters for input to the program, Prime (Genetics Computer Group; Madison, WI): T_m (from 70-60°C), G/C composition (from 55-40%), and minimum interval length (from 90%-80% of the length of the single copy interval).

Probe generation and chromosomal in situ hybridization.

DNA fragments were amplified by long PCR (Cheng et al., *Proc. Natl. Acad. Sci. U.S.A.*, **91**:5695-5699 (1994)) with LA-Taq as recommended by the manufacturer (Panvera, Madison, WI). Other enzymes for long PCR have been demonstrated to produce comparable results, including those manufactured by Roche Molecular Biochemicals, Indianapolis, IN; Stratagene, LaJolla, CA; and Invitrogen, Carlsbad, CA. The amplicons were purified by low-melt temperature agarose gel electrophoresis, followed by chromatography with Micro-con 100 columns (Millipore, Bedford MA), which removed contaminating extension products containing repetitive sequences.

Probe fragments were labeled by nick translation using modified nucleotides such as digoxigenin-dUTP or biotin-dUTP (Roche Molecular Biochemicals, Indianapolis, IN). Labeled probes were denatured and hybridized to fixed chromosomal preparations on microscope slides using previously described conditions (Knoll and Lichter, *Current Protocols in Human Genetics*, Vol. 1, Unit 4.3, Green-Wiley, New York (1994)), with the exception that preannealing of the probe(s) with repetitive DNA (such as C_ot1 DNA) was not utilized in a parallel set of hybridizations. Probes from a single chromosome region of ~100kb were hybridized individually or in combination to remove non-specific binding. Post-hybridization washes were performed at 42°C in 50% formamide in 2xSSC, followed by an additional wash at 39°C 2xSSC and one in 1xSSC at room temperature. Wash stringency was increased, if necessary, to remove hybridization of probes to related sequences elsewhere in the genome. Hybridized probes were detected

with a fluorochrome (such as rhodamine or fluorescein) tagged antibody to the modified nucleotide. Chromosome identification was performed by counterstaining the cellular DNA with 4', 6-diamidino-2-phenylindole (DAPI). Hybridized chromosomes were viewed with an epifluorescence microscope (Olympus, Melville, NY) equipped with a motorized multi-excitation fluorochrome filter wheel. Hybridization patterns on at least 20 metaphases (and 50-100 nuclei) were scored for each probe or combination of probes, with and without preannealing to C_0t1 DNA. Cells were imaged using a CCD camera (Cohu, Inc, San Diego, CA) and CytoVision ChromoFluor software (Applied Imaging, Santa Clara, CA).

Table 4 sets forth the data generated using the foregoing procedure, with respect to a number of probes specific to known disorders and cytogenetic abnormalities. The abnormality designation makes use of standardized nomenclature as set forth in ISCN 1995, An International System for Human Cytogenetic Nomenclature (1995), Mittelman F, ed.

Table 4

Disorder	Representative Cytogenetic Abnormality Detected by Probes on Metaphase Chromosomes	Gene or Transcript	Interval	GenBank Accession No.	Forward PCR Primer Coordinate, Beginning/End	Reverse PCR Primer Coordinate, Beginning/End	Sequence ID Forward/ Reverse Primers
Angelman Syndrome and Prader-Willi Syndrome	ish del(15)(q11.2q11.2)(UBE3A-)	*UBE3A*	IVS 8-IVS 9	AC004600	41085/41119	45354/45325	480/481
	ish del(15)(q11.2q11.2)(IC/ SNRPN-)	*IC/SNRPN*	IVS 3' to Exon u1B*	AC004737	13740/13769	15414/15387	482/483
Duplication 15 Syndromes	ish dup(15)(q11.2q13)(UBE3A++, IC/SNRPN++) and ish dic(15q11.2q13) (UBE3A++,IC/SNRPN++)	*IC/SNRPN*	IVS 5'-Exon u1B[1]-IVS 3'	AC004737	31102/31128	33347/33323	484/485
		IC/SNRPN	IVS 5'-Exon u1B	AC004737	47792/47821	49470/49441	486/487
Kallmann Syndrome	ish del(X)(p22.3)(KAL1-)	*KAL1*	~53 kb downstream	AC006062	104433/104465	107097/107072	488/489
		KAL1	IVS 10	AC006062	38822/38852	42042/42012	490/491
Turner Syndrome and Distal Xp Deletions	ish del(X)(p22.3)(SHOX-)	*SHOX*	IVS 1-Exon 2	NT_001151	44615/44646	47505/47473	492/493
		SHOX	IVS 2	NT_001151	49637/49669	52251/52217	494/495
		SHOX	IVS 3	NT_001151	54357/54387	56821/56791	496/497
	ish del(X)(p22.3)(GS2-)	*GS2*	Promoter-IVS 2	NT_001457	78970/79000	82994/82960	498/499
	ish del(X)(p22.3)(TBL1-)	*TBL1*	IVS 2	NT_001159	175379/175409	179665/179633	500/501
		TBL1	3' UTR	NT_001159	247264/247293	251290/251257	502/503
Down Syndrome	ish (21)(q22.2q22.3)(DSCR4x3)	*DSCR4*[2]	~39 kb upstream	AP000160	31007/31041	32999/32965	504/505
		DSCR4[2]	~30 kb upstream	AP000160	40725/40754	43078/43045	506/507
		DSCR4[2]	~20 kb upstream	AP000160	49973/50006	52409/52376	508/509
Chronic Myelogenous Leukemia and Acute Lymphoblastic Leukemia	ish t(9;22)(q34;q11.2)(BCR st)	*BCR*	Proximal to the major breakpoint in CML	U07000	100745/100776	104145/104115	510/511
		BCR	Proximal to the major breakpoint in CML	U07000	74699/74728	77903/77874	512/513
		BCR	IVS 8	U07000	114946/114978	117457/117426	514/515
	ish t(9;22)(q34;q11.2)(ABL st)	*ABL1*	Exon 1B-IVS 1B	U07561	27182/27213	29388/ 29357	516/517
		ABL1[3]	IVS 1B	U07562	9193/9222	11035/11004	518/519
	ish t(9;22)(q34;q11.2)(ABL mv)	*ABL1*	IVS 4-IVS 6	U07563	65951/65985	70266/70237	520/521
		ABL1	Exon 11-IVS 11	U07563	78862/78891	83813/83784	522/523

[1] u1B is ~160 kb upstream from the PWS shortest region of overlap and ~85 kb upstream from the AS shortest region of overlap.

[2] Probe also cross-hybridizes to a sequence found on the p arm of acrocentric chromosomes. Probe sequence is not found in public repetitive sequence database.

[3] Probe also cross-hybridizes to an interspersed repetitive sequence family that is not found in the public repetitive sequence database.

Disorder	Representative Cytogenetic Abnormality Detected by Probes on Metaphase Chromosomes	Gene or Transcript	Interval	GenBank Accession No.	Forward PCR Primer Coordinate, Beginning/End	Reverse PCR Primer Coordinate, Beginning/End	Sequence ID Forward/ Reverse Primers
Chronic Myeloge-nous Leukemia, Acute Lymphoblastic Leukemia	ish t(9;22)(q34;q11.2)(ABL mv)	*ABL1*	IVS 3	U07563	55807/55836	58077/58046	524/525
		ABL1[3]	IVS 3	U07563	53570/53604	55489/55455	526/527
		ABL1	IVS 3	U07563	55854/55888	57848/57817	528/529
		ABL1	IVS 4-IVS 6	U07563	66333/66367	70295/70264	530/531
Williams Syndrome	ish del(7)(q11.23q11.23)(LIMK1-)	*LIMK1*	IVS 13-3'UTR	NT_000398	59947/59976	62211/62187	532/533
		LIMK1	IVS 2	NT_000398	31966/31993	35015/34989	534/535
Acute Myelogcnous Leukemia- Type M4	ish inv(16)(p13q22)(PM5 sp)	*PM5*[2,4]	~20 kb downstream	NT_000691	24509/24538	27988/27958	536/537
		PM5[2,4]	~60 kb downstream	NT_000691	64204/64233	67682/67652	538/539
	ish inv(16)(p13q22)(PLA2G10 mv, PKD mv, PM5 sp)	*PLA2G10*[4] *PKD* *PM5*	IVS 3 IVS 12-Exon 15 ~100 kb upstream	NT_000691	68271/68300	71986/71957	540/541
		PLA2G10[4] *PKD* *PM5*	IVS 3 Exon 15-IVS 20 ~100 kb upstream & ~300 kb downstream	NT_000691	71957/71986	75481/75452	542/543
	ish inv (16)(p13q22)(ABCC1st)	*ABCC1*	IVS 6	NT_025903	313783/313812	315675/315645	544/545
Rubinstein-Taybi Syndrome	ish del (16)(p13.3)(CREBBP-)	*CREBBP*	IVS 18	NT_000671	58833/58862	63347/63318	546/547
Acute Lymphocytic Leukemia	ish t(12;21)(p13.2; q22.1)(AML1 st)	*AML1*	IVS 1-IVS 2	AP000057	98712/98741	102903/102872	548/549
	ish t(12:21)(p13.2;q22.1)(TEL/ETV6 mv)	*TEL/ETV6*	IVS 4	NT_000601	95456/95480	97283/97260	550/551
		TEL/ETV6	IVS 3	NT_000601	72543/72564	74385/74361	552/553
		TEL/ETV6	IVS 2	NT_000601	38216/38245	40091/40062	554/555
Cri-du-Chat Syndrome	ish del (5)(p15.3)(CTNND2-)	*CTNND2*	IVS 17	NT_000149	169655/169685	171976/171945	556/557
		CTNND2	IVS 14	NT_000149	199168/199202	203507/203473	558/559
	ish del (5)(p15.3)(SEMA5A-)	*SEMA5A*	IVS 3	NT_000147	23905/23935	27710/27676	560/561
		SEMA5A	IVS 3	NT_000147	30757/30790	33241/33209	562/563
		SEMA5A	IVS 3	NT_000147	14716/14748	17787/17753	564/565
	ish del (5)(p15.3)(SLC6A3-)	*SLC6A3*	IVS 3	AF119117	28206/28239	31894/31860	566/567
Langer-Giedeon Syndrome	ish del(8)(q23.3q24.1)(TRPS1-)	*TRPS1*	IVS 1	NT_002886	267731/267760	270758/270724	568/569
		TRPS1	IVS 1	NT_002886	271242/271271	274437/274404	570/571
Smith-Magenis Syndrome	ish del (17)(p11.2p11.2)(ADORA2B-)	*ADORA2B*[5]	Promoter-IVS 1	NT_000770	56443/56472	58524/58491	572/573
		ADORA2B[5]	IVS 1	NT_000770	77442/77475	79222/79189	574/575
	ish del (17)(p11.2p11.2)(FLI1-)	*FLI1*	IVS 9-IVS 12	U80184	6094/6127	7300/7267	576/577

[4] PM5 is ~1.3 mb telomeric of MYH11 gene, which is disrupted at the inv(16p) breakpoint. PLA2G10 is ~200 kb telomeric of PM5.

Disorder	Representative Cytogenetic Abnormality Detected by Probes on Metaphase Chromosomes	Gene or Transcript	Interval	GenBank Accession No.	Forward PCR Primer Coordinate, Beginning/ End	Reverse PCR Primer Coordinate, Beginning/End	Sequence ID Forward/ Reverse Primers
Smith-Magenis Syndrome (cont'd)		*FLI1*	IVS 12-IVS 14	U80184	7424/7453	8742/8708	578/579
		FLI1	IVS 15-Exon 21	U80184	9615/9647	11738/11704	580/581
	ish del(17)(p11.2p11.2)(MFAP4-)	*MFAP4*	IVS 2-3' UTR	NT_000760	132621/132654	134663/134634	582/583
	ish del(17)(p11.2p11.2)(ZNF179/PAIP1/ SHMT-)	*ZNF179-PAIP1*[3] *SHMT1*	Between ZNF179 and PAIP1. IVS 4	AL035367	9818/9850	12272/12241	584/585
	ish del(17)(p11.2p11.2)(LGLI/HUGL-)	*LLGL* *HUGL*	Promoter –Exon 2 Promoter – IVS1	AL035367	1194/1226	5365/5334	586/587
Charcot-Marie –Tooth Disease Type 1A	ish dup(17)(p11.2p11.2)(PMP22++)	*PMP22*	Promoter (~5 kb upstream)	AC005703	153173/153202	155027/154994	588/589
		PMP22	~22kb downstream	AC005703	215632/215661	217362/217329	590/591
		PMP22[5]	IVS 3	AC005703	184666/184700	186035/186006	592/593
		PMP22	IVS 3	AC005703	176746/176778	179073/179044	594/595
Miller-Dieker Syndrome	ish del(17)(p13.3)(PAFAH1B1/EIF-3-)	*PAFAH1B1*	~5kb downstream	NT_000774	63645/63679	66603/66573	596/597
		EIF-3[6]	IVS 24–IVS 27				
		PAFAH1B1	~7-8 kb downstream	NT_000774	68841/68870	71195/71163	598/599
		EIF-3	IVS 15-IVS 19				
		PAFAH1B1	~13 kb downstream	NT_000774	75328/75362	78122/78093	600/601
		EIF-3	IVS 5-IVS 11				
Alagille Syndrome	ish del(20)(p12.3p12.3)(JAG1-)	*JAG1*	IVS 5-IVS 8	AL035456.24	153935/153966	157675/157642	602/603
		JAG1	IVS 2-IVS 3	AL035456.24	144875/144904	147028/146995	604/605
		JAG1	Exon 1-IVS 2	AL035456.24	135644/135673	139440/139407	606/607
Monosomy 13q32	ish del(13)(q32.3)(ZIC2-)	*ZIC2*	~5.8 kb downstream	AL355338	111114/111145	116046/116012	608/609
Trisomy 13	ish (13)(q32.3)(ZIC2x3)	*ZIC2*	~2 kb upstream	AL355338	128595/ 128627	133039/133006	610/611
Wolf-Hirschorn Syndrome	ish del (4)(p16.3)(HD-)	*HD*	Exon 67	NT_000102	267614/267643	271120/271091	612/613

[5] Probe was hybridized in combination with other probes and not individually.

[6] Probe is downstream and adjacent to PAFAH1B1 gene (formerly known as LIS1). An expressed transcript homologous to EIF-3 is found at these coordinates.

Example 4

In this example, a more precise chromosomal breakpoint determination was made using the probes of the invention. Structural chromosome rearrangements can be inherited in genetic disease or acquired as in the case of certain cancers. They can occur within a single chromosome (such as an inversion, deletion or duplication) or between homologous or non-homologous chromosomes (i.e. translocations). With the probes of the invention, the precise region of breakage can be determined at a previously unprecedented level of resolution. Such resolution permits detection of genes or sequences that are disrupted in the formation of the rearrangement and may provide insight into etiology, prognosis and/or treatment. In inherited contiguous gene syndromes, precise localization of chromosome breakpoints can define the extent of deletion or duplication and hence the prognosis of the disorder (Cheng et al., Am. J. Hum. Genet., **55**:753-759 (1994)).

The following example illustrates how the single copy probes hereof provide more precise information than commercially available cloned probes for the same chromosomal region.

In most cases of chronic adult myeloid leukemia (CML; 90%) and in some cases of acute lymphoblastic leukemia (ALL; 25-30% adults; 2-10% children) (Perkins et al., *Cancer Genet. Cytogenet.*, **96**(1): 64-80 (1997); Rubintz et al., *J. Pediatr. Hematol. Oncol.*, **20 (1)**:1-11 (1998)), a reciprocal translocation between chromosome 9q34 and 22q11.2 is evident (Rowley, *Nature*, **243**:290-392 (1973)). The abnormal or derivative chromosome 22 that results from this translocation fuses the ABL1 oncogene on chromosome 9 to the BCR (breakage cluster region) promoter on chromosome 22. The ABL1 oncogene is expressed as either a 6 or a 7 kb mRNA transcript with alternatively spliced first exons, exons 1b and 1a respectively, spliced to the common exons 2-11. Exon 1b is ~250 kb proximal of exon 1a and this very long intron is a primary target for translocations. In CML, the ABL1 gene is translocated from chromosome 9 to the promoter of the BCR gene on chromosome 22 to produce a chimeric BCR-ABL1 protein (Bernards, et al., *Mol. Cell. Biol.*, 7:3231-3236; (1987). The BCR gene contains 24 exons and encodes a 160kD protein. The BCR breakpoints differ in CML and ALL. In CML, most breakpoints occur within the 5.8 kb major breakpoint cluster region (M-

BCR) which corresponds to exons 12 through 17 (or b1 through b5); whereas in most ALL, the BCR gene breaks between exons 1 and 2 (minor or m-BCR). Thus different molecular rearrangements resulting in differently-sized proteins occur in each disorder. These rearrangements can be distinguished with the probes of the invention.

At the DNA level, dual color-dual probe fluorescence in situ hybridization (FISH) strategies are available to detect the BCR-ABL1 translocation on metaphase chromosomes and interphase nuclei (Bentz et al., *Blood*, **83**:1922-1928 (1994); Sinclair et al., *Blood*, **90**:1395-1402 (1997); Buno et al., *Blood*, **92**:2315-2321 (1998)). With conventional FISH, the initial strategy was to have the probe for the ABL1 oncogene region (distal of the breakpoint and ~200kb in size) detected in one color (i.e., red) and the BCR region (proximal of the breakpoint) in another color (i.e., green). Thus, in derivative chromosome 22 positive cells, one red and one green signal co-localized to give a yellow hybridization signal indicating the presence of a derivative 22 chromosome while the normal chromosome 9 and chromosome 22 remained as independent red and green signals. More recently, larger cloned DNA probes that span both sides of the ABL1 translocation breakpoint have been utilized for detecting translocations. In this strategy, the part of the probe that is proximal to the breakpoint remains on the abnormal chromosome 9 and the part distal to the breakpoint co-localizes with BCR as in the previous strategy. This results in an extra signal (ES) on the translocated chromosome 9 and is the strategy behind the BCR-ABL1 ES dual-color translocation probe (Vysis, Inc., Downers Grove, IL) that many clinical cytogenetics laboratories use (Herens et al., *Br. J. Haem.*, **110(1)**:214-216 (2000); Sinclair et al, *Blood*, **95**(3):738-744 (2000)). In the ES system, the ABL1 probe cocktail spans a genomic target significantly larger than the ABL1 gene. This cocktail extends from arginosuccinate synthetase gene (ASS), which is ~250 kb upstream from ABL1, through the ABL1 gene and several kb downstream.

Several single copy probes were designed for ABL1 and BCR by identifying the boundaries of single copy intervals at these loci. Figs. 13 and 14 indicate the locations of potential single copy probes within the BCR and ABL1 loci, respectively. Eleven intervals exceeding 2 kb length are distributed throughout the BCR gene, 5 of which have been currently designed as probes. A similar number of additional shorter single

copy regions in this gene (between 1.5 and 2.0 kb) could be combined to more precisely delineate translocation breakpoints. In the ABL1 gene, 10 intervals > 2kb are found, six of which have been designed as probes.

Multiple single copy probes for both BCR (Fig. 13) and ABL1 (Fig. 14) have been deduced and oligonucleotide primer sets derived. For BCR, these probes discriminate between the minor (Fig. 13) and major breakpoints (Fig. 13 and SEQ ID Nos. 510-515). Conventional FISH testing with BCR-ABL1 ES probe does not distinguish between rearrangements at the major (M) and minor (m) breakpoints in the BCR gene. A mixture of m-BCR probes detects the minor breakpoint often seen in patients with ALL, and those designated M-BCR (SEQ ID Nos. 510-515) detect the major breakpoint evident in most patients with CML. If all of the probes translocate to the chromosome 9, then the gene is interrupted at the minor breakpoint. If only the M-BCR probes translocate from chromosome 22 to the derivative chromosome 9 and the m-BCR probes remain on the derivative chromosome 22, then the gene is interrupted at the major breakpoint.

For ABL1, two of the probes are predicted to be proximal of the breakpoint (SEQ ID Nos. 516/517 and 518/519), and the others are distal to the breakpoint (SEQ ID Nos. 520 through 531). Hybridization of three ABL1 probes distal to the breakpoint in CML shows that the probes have moved to the derivative chromosome 22 (Fig. 15), whereas a combination of five ABL1 probes that span the breakage interval demonstrates that some probes remain on the derivative chromosome 9 while others move to the derivative chromosome 22 (Fig. 16). Based on these results and the positions of these probes, it can be deduced that the breakpoint interval spans positions 11004 bp through 65951 bp of Fig. 14. It will be evident to one skilled in the art that the region of breakage can be more precisely refined by hybridizing probes from the single copy intervals between Seq. ID Nos. 519 and 520. The exact location at the breakpoint can be then determined from the genomic sequence of the refined breakage region.

Recent studies (Herens et al, *Br. J. Haem.*, **110**:214-216 (2000)) utilizing the commercially available ES probe have demonstrated that ~10% of CML patients do not have an extra hybridization signal on the derivative chromosome 9 because sequences

are deleted upstream of the ABL1 gene. In some instances, the deletions extend as far as the ASS gene and such a deletion is associated with poor prognosis (e.g., blast crisis). These large FISH probes are not useful for detecting interstitial deletions in the region between ASS and ABL1, as evidenced by an increased deletion detection rate of up to 1/3 of CML patients when shorter probes of ~100-200 kb are hybridized (Sinclair et al., *Blood*, **95**(3):738-744(2000)). Since some patients harbor deletions of sequences proximal to the ABL1 breakpoint on the derivative or translocated chromosome 9, single copy probes are being used to delineate the extent of hemizygosity in this chromosomal region. Correlation of deletion breakpoints with clinical outcomes will determine if the loss of specific genes in this chromosonal interval is prognostic for clinical findings such as early blast crisis.

Example 5

This example demonstrates that increased hybridization signals can be generated using probe sequences from duplicated genomic domains. Several single copy probe sequences were designed, which were a part of highly similar duplicon or triplicon domains as previously described.

Several probes from chromosome 16p13.1 (SEQ ID Nos. 536-543), close to the inversion breakpoint in Acute Myelogenous Leukemia – Type M4, contain sequences that are a near perfect triplication of sequences in this region. In the genome draft sequence, two of these domains are tandemly arranged, separating the probe sequences by 40 kb, and a third telomeric interval is separated by 1.2 mb. The sequences of the three intervals differ by only ~1.5%. Hybridization with this probe demonstrated two clustered, but clearly separable signals. One hybridization corresponds to the combined first and second paralogs and the other to the third copy of this sequence.

A probe from the chromosome 17p11.2 interval that is commonly deleted in Smith-Magenis Sydrome (SEQ ID Nos. 586-587) contained a near-perfect triplication. The probe was intended to detect a deletion within a near single-copy sequence in IVS4 of the SHMT1 gene. However, two paralogous subsequences separated by ~15 kb, exhibiting 99.8% identity with the SMHT1 sequence, were also detected in the genome draft between the ZNF127 and PAIP1 genes ~2.7 mb centromeric but also within the

genetic interval commonly deleted in patients with this disorder. Due to the proximity of these sequences to the chromosome 17 centromere (which is a highly condensed region), a single hybridization locus was observed.

Two probes from the Down syndrome critical region (SEQ ID Nos. 504/505-508/509) were each embedded in the same large duplicon, which was ≥80 kb in length. These duplicated sequences are separated by 1.1 mb, and reside, respectively at the centromeric and telomeric ends of chromosome band 21q22.2. Despite the fact that the duplicated sequences were separated by 1.1 mb, a single hybridization signal was detected in this region of chromosome 21 using each probe. Therefore, this region of chromosome 21, like chromosome 17p11.2, seems more condensed in metaphase chromosomes than sequences in 16p13.1, in which duplicated regions separated by similar distances were distinguishable.

Single copy probes developed from such regions, of course, lack known repetitive sequence elements. However the probes generally hybridize to all of the paralogous copies, since each of the copies remain hybridized even under the most stringent hybridization wash conditions. Because multiple, tightly clustered sites on the chromosome are hybridized in a specific interval, the hybridization signal produced from these hybridizations is brighter than that expected from a comparable probe sequence which was represented once per haploid genome. Thus, these genomic duplicons or triplicons increase the effective target size of the probe. This implies that shorter probes from such regions can produce hybridization signals comparable in intensity to those generated by longer probes. Selection of shorter probes from duplicated genomic domains will be particularly useful for development of probes for genomic regions where long single copy intervals are underrepresented.

Example 6

The increasing availability of accurate draft human genome sequences has facilitated development of single copy probes in accordance with the invention for many previously inaccessible chromosomal regions. Although the most current comprehensive up-to-date sequence databases have been used to detect repetitive elements (http://www.girinst.org/repbase) present in these draft sequences,

hybridization of single copy probes to metaphase chromosomes has revealed that several probes contain previously unrecognized repetitive sequences. This was determined by documenting hybridization of a probe to the homologous chromosomal band where it is known to be mapped as well as other locations not found in the draft genome sequence.

The draft genome sequence is incomplete, with ~90% of the euchromatic genome having been sequenced (International Human Genome Sequencing Consortium, Initial Sequencing and Analysis of the Human Genome, *Nature*, **409**:860-922, (2001). It was anticipated that some repetitive sequence families, especially those present among the missing sequences, would have not been detected. Despite screening for known repetitive sequences, several euchromatic single copy probes appear to contain homologs of repetitive sequence families that are predominantly found in multiple copies on the short arms of acrocentric chromosomes (chromosomes 13, 14, 15, 21, and 22). These probes included three sequences derived from the Down Syndrome critical region on chromosome 21 (amplified with SEQ ID Nos. 504/505, 506/507 and 508/509), two sequences from chromosome 16p13.1 that straddle the site of chromosomal inversion in Acute Myelogenous Leukemia, Type M4 (amplified with SEQ ID Nos. 536/537 and 538/539).

Such chromosomal domains, termed nucleolar organizer regions, are known to contain thousands of copies of the ribosomal RNA cistrons arranged in long tandem arrays (Sylvester et al., Hum. Genet.,; **73**:193-8 (1986)). The human genome draft genome sequence is devoid of contiguous sequences from these chromosomal regions. The International Sequencing Consortium eliminated clones containing these and other tandemly repeated sequences (e.g., heterochromatin) from consideration early in the sequencing effort, since it was recognized that any assembly of sequences from such clones would be ambiguous, and thus unreliable. Other sequences distinct from, but co-localizing with, ribosomal RNA genes would most likely also be tandemly arranged in chromosomal nucleolar organizer regions. Because of the lack of sequence information from these intervals, the distribution of sequences within them that are related to sequences elsewhere in the genome has not been previously appreciated. While single copy FISH with these probes demonstrated localization to the expected euchromatic

intervals, additional significant hybridization to the short arms of several acrocentric chromosomes was seen. Addition of C_0t 1 DNA prior to hybridization removed cross-hybridization to these repetitive sequences. These additional signals are consistent with tandemly organized multiple copies of sequences related to the probe on short arms of acrocentric chromosomes.

In addition, hybridization of presumed single copy probes to interspersed repetitive sequence families was detected, despite the fact that these probes were filtered for repetitive sequences using RepeatMasker software. One probe, mapping to chromosome 17p11.2, within the interval commonly deleted in Smith-Magenis syndrome (amplified with SEQ ID No. 586/587), was found to cross-hybridize with interspersed repeats. Sequences close to the translocation breakpoint at chromosome 9q34 (amplified with SEQ ID Nos. 518/519 and 526/527) also potentially contain interspersed repetitive sequences, based on hybridization of combinations of these probes with another probe from this region. The hybridization signals at the mapped locations for these probes were not stronger than those observed for the cross-hybridizing sequences, nor were the cross-hybridizing sequences removed by increasing the stringency during the washing procedures. This suggests that the probes contain previously unrecognized repetitive sequence families, rather than highly divergent copies of known interspersed repeats (whose failure to be recognized by RepeatMasker software would have led to their inadvertent inclusion in the designed probe).

Although all of these probes appear to contain members of previously unrecognized repetitive sequences families, the probes themselves are likely to be composed of both single copy and repetitive sequences. It is feasible to separate these sequence components by iterative hybridization of different PCR-generated subsets of each probe sequence to chromosomal DNA. However, since each entire probe sequence is known, the sequence can be added to the repeat sequence database used to develop new, additional single copy probes. The probe sequences are not lengthy (some interspersed repeat families are, in fact, longer, e.g., L1, than the longest single copy probe); therefore, minimal additional computational overhead is incurred by addition of these sequences to the database of human repetitive sequence families. The addition of these previously unknown repetitive sequence family members results in a more

comprehensive repetitive sequence database, that in turn improves the design of single copy probes. Single copy probes subsequently designed using the larger repeat sequence database, will not contain these novel repetitive sequences. This heuristic algorithm improves the purity of single copy sequences in single copy probes.

Generally speaking, the invention thus provides a method of determining the existence of previously unknown repeat sequence families in a genome. This method involves reacting a labeled, putative single copy nucleic acid probe with the genome, and causing the probe to hybridize. If the probe hybridizes at more than three different locations (and preferably at more than ten different locations), then it is likely that a new, previously unknown repeat sequence has been found.

All references cited above are expressly incorporated by reference herein. In addition, the subject matter of Disclosure Document #471449 filed March 27, 2000, is also incorporated by reference herein.

We claim:

1. A nucleic acid hybridization probe comprising a labeled, single copy nucleic acid which will hybridize to a deduced single copy sequence interval in target nucleic acid of known sequence, said nucleic acid probe having a length of at least about 50 nucleotides.

2. The probe of claim 1, said probe including a plurality of different, labeled nucleic acids each of which will hybridize to respective deduced single copy sequence intervals in said target nucleic acid, each of said nucleic acid probes having a length of at least about 50 nucleotides.

3. The probe of claim 1, said nucleic acid probe having a length of at least 100 nucleotides.

4. The probe of claim 3, said nucleic acid probe having a length of at least about 2000 nucleotides.

5. The probe of claim 1, said target nucleic acid being selected from the group consisting of DNA, RNA and mRNA.

6. The probe of claim 5, said target nucleic acid being DNA.

7. The probe of claim 1, said nucleic acid probe being single stranded.

8. The probe of claim 1, said probe being essentially free of blocking nucleic acid sequences which will hybridize to repeat sequences within the genome of which said target nucleic acid is a part.